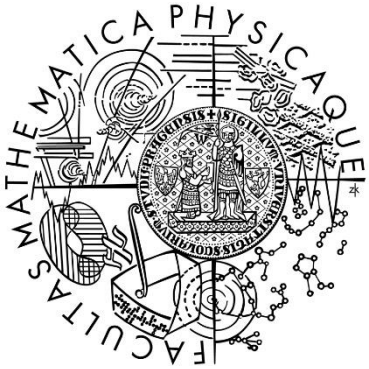
Charles University in Prague

Faculty of Mathematics and Physics

MASTER THESIS



Andrej Kalický

High Performance Analytics

Department of Software Engineering

Supervisor of the master thesis: Ing. Vladimír Kyjonka

Study programme: Software Systems

Specialization: Database Systems

Prague 2013

Herewith, I would like to thank my supervisor for his dedicated time, inspirations, thoughts and consultations and SAS Institute that provided the software and analytical platform for an experimental assignment. Further, I would like to thank my family and friends for their support along the way of accomplishing this project.

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In ..... date ..... signature

Název práce: Vysoce výkonné analýzy

Autor: Andrej Kalický

Katedra / Ústav: Katedra softwarového inženýrství

Vedoucí diplomové práce: Ing. Vladimír Kyjonka, Katedra softwarového inženýrství - externí pracovník, SAS Institute ČR, s.r.o.

Abstrakt: Tato práce se zabývá fenoménem velkých dat, charakterizovaných rychlým nárůstem jejich velikosti, variability a dynamiky, v souvislosti s posunem paradigmatu jejich zpracování. Cílem této práce je poskytnout přehled a nadlhled nad celou problematikou a podat ucelený a konzistentní obraz v oblasti High Performance Analytics včetně problému a výzev velmi rychle se rozvíjejícího pokročilého analytického zpracování. Přehled HPA přináší shrnutí a utřídění, popsání výhod a přínosů jednotlivých metod HPA, které využívají různou kombinaci systémových prostředků. Praktická čast této práce je zaměřená na realizaci úlohy analytického zpracování velkého vzorku dat vybranými metodami s využitím softwarové platformy SAS Institute. Na tomto příkladu je ilustrováno uplatnění metody in-memory analýzy s vyhodnocením výkonu a vhodnosti použití této metody na analytickych príkladech a operacích.

Klíčová slova: Velká Data, pokročilé analýzy, in-memory analýzy, in-database analýzy

Title: High Performance Analytics

Author: Andrej Kalický

Department / Institute: Department of Software Engineering

Supervisor of the master thesis: Ing. Vladimír Kyjonka, Department of Software Engineering - external specialist, SAS Institute Czech Republic, s.r.o.

Abstract: This thesis explains Big Data Phenomenon, which is characterised by rapid growth of volume, variety and velocity of data - information assets, and thrives the paradigm shift in analytical data processing. Thesis aims to provide summary and overview with complete and consistent image about the area of High Performance Analytics (HPA), including problems and challenges on the pioneering state-of-art of advanced analytics. Overview of HPA introduces classification, characteristics and advantages of specific HPA method utilising the various combination of system resources. In the practical part of the thesis the experimental assignment focuses on analytical processing of large dataset using analytical platform from SAS Institute. The experiment demonstrates the convenience and benefits of In-Memory Analytics (specific HPA method) by evaluating the performance of different analytical scenarios and operations.

Keywords: Big Data, advanced analytics, in-memory analytics, in-database analytics

Contents

1 Introduction

The world has turned into information society that highly relies on data. Since information systems generate enormous amounts of records every day, every second, it seems the world is reaching the level of data overload. It is obvious now, that in order to process such volumes of data an enormous capacity is required in terms of storage and computing resources. Whereas the growth of capacity is limited by evolution of hardware and technologies, the growth of the data volume is in fact unlimited.

Getting more specific, nowadays many organisations has adopted and broadly use information systems running on technological platforms, many their agendas has become addicted to data. In mature organisations data directly affect the logic of business processes, information has become a core of their business or business end. Hence business demands the data, furthermore availability of specific data in specific time. More and more complex and risky decision making process relies on correctness and transparency of data.

1.1 Motivation

Interesting driver related to this topic mentions that the growth of data is unlimited. What is the society going to do about the data overload? How to handle and moreover to process all the data? Seems like we are having the Big Data issue.

Another driver for this topic is retrieving the information (not to gather all data for further analysis). Among all the data, how to retrieve the relevant information and within a required time? Which analytics should be applied on data? What is the balance between cost of retrieval and value of that information? What are the costs of capacity to retrieve desired information? It seems like it is all about the profit, trade-off between value of information and the cost to get it. Additionally to both drivers the challenge is to visualise the information in such a way that its value is comprehensive and understandable. The main issue is the information overload.

Analytics in the traditional mode, in terms of the Big Data, are acquiring data that may or may not be needed for analysis. This all requires an innovative point of view, a different approach, architecture or infrastructure, if any. High performance analytics is one of them.

Adopting new technologies requires to process, discover and analyse these massive data sets that cannot be dealt with using traditional databases and architectures due to the lack of capacity resources in terms of computation and storage. High performance analytics represents one of the innovative approaches that can be applied on the increasing volumes, velocity and variety of data.

1.2 Goals

Big Data Phenomenon, which is characterised by rapid growth of volume, variety and velocity of data - information assets, thrives the paradigm shift in analytical data processing. High Performance Analytics (HPA) can be considered as one of the approaches. The aim of the thesis is a research (overview, classification, discussions on problems and challenges) on the pioneering state-of-art of advanced analytics utilising various methods (HPA methods) that could escalate and optimise the computation performance of analytics.

Considering the fact that the selected area of research is currently being refined and formalised and simultaneously is emerging rapidly in proprietary definitions and solutions from multiple vendors, the goal of the thesis is to classify and provide summary and overview

with complete and consistent image about the area of High Performance Analytics. Moreover, utilisation of these methods shall be demonstrated in practical assignment involving a processing of huge dataset.

1.3 Outcome

The scope of the thesis is dedicated to research and approaches of Big Data and High Performance Analytics. Theoretical part of the thesis is an outcome of comprehensive research that summarises a state-of-art overview for this problem, defines the drivers and consequences of Big Data Phenomenon, and introduces approaches for handling Big Data, in particular approach based on High Performance Analytics.

Specifically the outcome of the research is oriented on an overview of HPA, classification, characteristics and advantages of specific method of HPA utilising the various combination of system resources.

Practical part of the thesis is an outcome of experimental assignment that includes analytical processing of large dataset using analytical platform from SAS Institute. The experiment demonstrates analytical processing for selected HPA methods that are discussed in theoretical part. One part of the experiment includes composing different analytical scenarios on which the advantages and convenience of HPA platform are demonstrated.

1.4 Structure

The reminder of this thesis is structured as follows. The thesis has two parts, theoretical and practical. The theoretical part consists with of the following sections.

Chapter 2 covers the identification of the problem from theory and business perspective. It summarises the reasons for data analysis, importance of Business Intelligence together with its downside in terms of Big Data, and opens the rationale for high performance analytics.

Chapter 3 introduces the phenomenon of Big Data that starts with causalities and definitions, continues with influence and impacts of Big Data on processes, infrastructure, architecture of data management and analytical system, and ends with approaches how to handle them. Chapter 4 covers the introduction of High Performance Analytics, definition, classification of HPA methods, emerging technologies and techniques for HPA realisation.

The practical part consists with of the following sections. Chapter 5 specifies the goal and scope of experimental assignment together with its limitations. Chapter 6 continues with description of environment (infrastructure, software, analytical platform), determines the process with analytical operations for the experimental assignment and defines evaluation technique for the experiments.

Chapter 7 documents the implementation of experimental assignment. In particular, it determines the size and content of datasets and objects used in implementation. It documents the modelled analytical scenarios and adds samples of data, metadata and outputs.

Chapter 8 unveils the results for particular analytical scenarios and analytical operations performed during experiments, evaluates results and discusses the findings, advantages and limitations of selected analytical platform on which experiments are held.

2 Problem Identification and Summary

As mentioned in introduction a central point of this thesis is data, data processing, extracting information and stuff around it. Let us first start with theoretical approach of problem.

2.1 Theoretical Problem

The data volume represents a challenge. Well, not just like that, should be placed within a context. In the work in [13], Gartners started with **available data** (customer data in their business context) that grows in all dimensions (dimensions are detailed in sections later), and they associated data with **analytical** and **execution capacity**, illustrated in Figure 1.

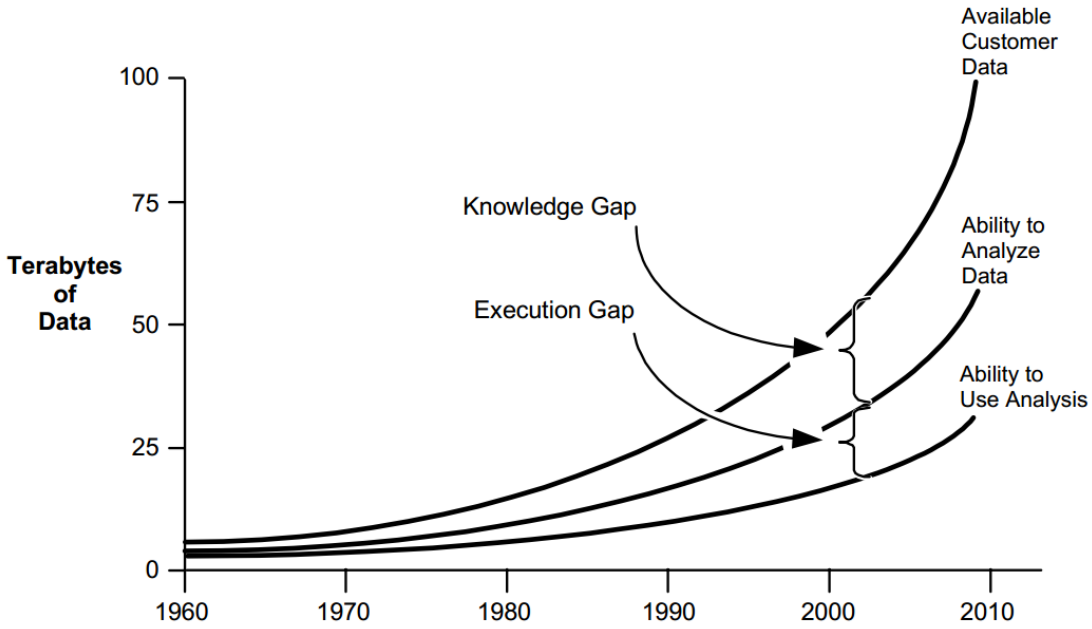


Figure 1: The data volume challenge. [13]

Considering the actual trends, when all of lines are growing, the availability of data has overloaded a capability to analyse data (analytics), as well as a capability to use the analysis – either to run analysis or store analysis (computing and storage capacity). A **knowledge gap** expresses the inability to analyse data due to the limited analytical techniques (might include data mining algorithms, natural-language processing, etc.), e.g. advanced sentiment analysis of textual comments of social media. An **execution gap** expresses inability to utilise analysis due to the limited availability of resources (might include processing units, data storages, etc.) for requested period of time, e.g. task to process the all daily transaction at bank in requested format of the clearing house over one night.

Where is the issue?

While an execution capacity in terms of hardware is growing at more or less stable rate, data volumes are growing exponentially. Therefore the knowledge gap is getting wider, as well as the area of the lost information opportunities in set of available data containing the information (relevant in respect to the information needs). The aim is to maximise the set of relevant data that potentially contains the valuable information. Therefore, the volume and availability of data are not a problem rather than the processing and organisation of data.

Approaches?

The problem illustrated in Figure 1 may have different resolving approaches. Hypothetically, the knowledge gap can be closed by limiting or reducing growth of data. Obviously, this is not going to happen. Analytical capacity is determined and dependent on research (advanced analytics). Well, having mature level of analytics might be a good approach unless the analytics are physically possible to run on disposal technologies and hardware. The execution gap can be potentially closed by increasing utilisation capacity, such as distributed or parallel processing (limited by level of segmentation of task) accompanying with additional units for processing (CPU, RAM) and storing data. This approach has downside in worldwide limited resources (example: BitCoint) and will not solve the problem due to high ratio supply:demand of data.

Apart from capacity another perspective of optimising may be input itself. Is all data needed? Necessary to analyse or store? Based on the requirements the redundant or irrelevant data (that holds small, if any, information) can be filtered out, with risk of lost information opportunity in missing data (yet we don't know what we don't know). By applying “brute-force” computation on whole dataset the problem is back to square one.

2.2 Business Case

After a few paragraphs that mentioned the theoretical problem, let’s try to bring up some ideas about how this problem might be reflected in real business environment. Following business case is chosen because it covers most of problem of data processing and management.

Getting Started

Imagine the mobile telecommunication operator (for simplicity of reference the fictive operator in this scoped is called BeyeMobile) that provides services (among others) calls, sms or mms, (mobile) internet, TV broadcasts, etc. All of those activities generate data, or their existence (in past) can be described by data. Let’s start simple: existence of one (instance of) activity is one record. Well, of course the reference to the client account should be kept, type of activity, timestamp (when it happen), optionally duration (if phone call, or data connection slot), data traffic (e.g. size of MMS, data traffic while using mobile internet, etc). Let’s call those records **primary data**.

What’s next? As BeyeMobile doing business, thus creating - in ideal case maximising - profit so the shareholders are happy. Monthly (or another defined period) invoices are sent out. Where did the invoices come from? Hmm. Pre-defined billing process creates them. This process refers to **secondary processing** that creates **secondary data** (not the least secondary data in this case). Users consume services (mobile activities) they pay for (or they pay for availability of these services), BeyeMobile is making money, everyone is happy, unless…

Unless there is open competitive market with mobile telecommunication services and users are free to choose the best solution or trade-off solution for the mobile activities they want to use. Users might not be satisfied with the price, availability, technical parameters of services, and start re-considering options on the market. What will happen in the situation described above?

Analytics?

Well in BeyeMobile, the profit and loss (PnL) is decreasing. How did our teclo company have found out? That’s where the Business Intelligence (BI) is coming in place and not only. For

simplicity and brief of this business case, some analytics exist - analytical methods exists that they can analyse data and find out or get some results. BeyeMobile is a smart company and it has adopted BI solution and established one key performance indicator, and that is, currently decreasing, PnL. That's how Beye founded out, by using **descriptive analytical method** that performed **secondary processing:** revenue minus expenses. Ceteris paribus, revenue is calculated from either primary data (records of activities) or secondary data (invoices), expenses of BeyeMobile coming from other data source that is different from provided services.

Being a CEO of BeyeMobile knowing the trend of decreasing PnL, one should ask immediately, why is it happening? Why? Clients' insolvency? Frauds? Client's migration? New competitor? Alternative offers, campaigns, products or bundles at competitors? Alternative technology in mobile or internet communication? Well, BI has **diagnostic analytics** for this issue that explore the reasons and causalities. Obviously, to provide the answers for new questions BI needs more additional or external data and sources to feed analytics. For instance to detect frauds, analytical model should contain pre-defined (trained) patterns that unveil them.

Advanced Analytics?

To continue our story, or business case, BeyeMobile at the end found out the causality of decreasing PnL trend. BeyeMobile might take out the immediate action, if there is any with doable effect within short timeframe for relatively high ROI – return on investment (action costs money, right?). Or one may ask question, what is going to happen with decreasing PnL trend? When the trend will get stabilised, if any? Will the trend continue decreasing linearly, or more intensely? What if the trend is stabilise in two month, likely the management will not take any expensive action. What if the trend is going down rapidly, with taken action what will be the return on investments both to cover the expenses of action and the losing profit as well?

Knowing the answers for these question will affect the decision on immediate action, which should be taken out in order to stabilise PnL trend. **Predictive analytics** from BI family are suitable for these sort of analysis (thoughtful reader may notice the "what if" used at the beginning implying what-if analysis).

Let's now add more data from other sources into our story of successful BeyeMobile. Data from social media (e.g. Facebook, Twitter, etc). Are our customer (and not only) complaining about BeyeMobile services? Are they trying to compare BeyeMobile's services with other competitors? What is the response on campaign through social media? Using **sentiment analysis** (from family of natural-language processing analyses) may provide the subjective opinion of their posts or comments in the world of social media, whether is positive or negative (for simplicity). From results of sentiment analysis BeyeMobile can predict trends as well.

Iteratively, the BI division of BeyeMobile may add more data with different structures into **decision making process**. Data generated in BeyeMobile's call center about complaints, claims, providing informative support, etc. All those data may help in predictions. Analytical tools can parse out all the necessary information that could provide hindsight or insights about issue with PnL trend.

Is the problem of PnL trend getting more complex or more complete? Let's make it harder. The BI division decided to extend analytics with **prescription analytics**, which support decisions by giving recommendations on possible actions that may lead into defined successful expectation (e.g. to improve PnL). In scenario that BeyeMobile has decrease in PnL and it is due to the rapid growth of frauds (reminder: that are detected by **diagnostic analysis**), the prescription analytics should be notified about this cause and with available knowledge of ("telco") business should be able to provide the feedback in preventing the frauds, at least with two output. Frist, how to constantly improve the model that detects frauds, and second how to recommendation about prospect modification into business model that frauds will be diminished. Sounds like sci-fi?

Imagine another reason of decreasing PnL such as clients' migration towards competitors or another technology. Indeed, it has been detected at BeyeMobile because of lower revenue, clients do less activities with their mobile phones, or using massively internet connection instead of sms or mms services, or bad economic situation nowadays? If the reason is client's migration, business manager at BeyeMobile expect recommendations in a sense, should focus marketing campaign on new acquired clients rather than on a client retention programme? Should explore the market and adjust the business model by providing new bundles, programmes, products, activities for clients? Or combination of recommendations?

Challenges?

By the way, all kind of analytics mentioned in successful case have been introduced as part of **secondary processing**. The term "secondary" may invoke the feelings that the analysis has some pre-conditional step. Innovative souls at BeyeMobile are perhaps trying to find out synergy form secondary to primary processing, to have the information faster, to know faster than competitors, to be first on market, ideally to have the information out of analytics in real-time, right? More accurate information?

One may observe that descriptive and diagnostic analytics are analysing the past data or current streams of data and trying to be explanatory. Others, predictive and prescriptive may be more uncertain, they look into the future. Why would BeyeMobile make investments in such an analytical system in volatile and competitive market? To make money.

2.3 Business Drivers

One of the objectives for business organisations is the maximisation of profit. Theoretically, that can be achieved by having the completive advantage. Holding the right information that others doesn't can differentiate the market leader and followers.

Organisations has become hungry for information, how to interpret it into the moment of surprise, to be first on the market while others haven't started thinking about the idea yet. As important as market entry is leaving the market, to know when is appropriate time for exit decisions (of product). Going further, knowledge of insights of market data can lead into discarding competition. Insights can be also interpreted into optimisation of business process itself. With implementing of this feedback loop companies thrive to established agile strategies (short term as well as long term).

Data analysis can research a market data. Investigate the customer's behaviour and segmentation of customers helps to suitably focus their campaigns. Campaigns represent investments for organisation that also require proper timing. Seeing the trends and

opportunities in timely manners, being able to move quickly when market moves, moreover move before the market moves, those all are concerns of organisations pursuing the profit.

Nowadays business analytics system (for instance performance management applications, analytic apps, business intelligence tools, data-warehouse management platforms) are being used in diverse areas of business: finance (for budgeting, planning, strategy), supply chain (for procurement, logistics, inventory), customer relationship management (in sales, customer service, contact centre, marketing, web site analytics, price optimisation), service operations (in banking, education, government, healthcare), reporting and analysis tools (for dashboards, OLAP, ad-hoc queries), advanced analytics (in data mining, statistics).

All operations and process above, seems like everything is named, are subject to optimisation (optimising costs, time, and revenue). That is all about business needs and concerns. Next section continues to unveil transformations of these needs into insights.

2.4 Insights to Foresights

Advanced Analytics can be applicable for various business analysis with regards to explore customer trends (behaviour, competition), fraud detection, inefficiency in business process (Capability Maturity Model - CMMI), market basket analysis (dependencies, causalities, relations in products' sales), etc. The bottom line is

Diverse analysis utilisations can be categorised according to the velocity of data with time dependencies (real-time, batch processing), or to the variety of data (structured, semi-structured, unstructured). Overview of use cases for analytics is depicted in Figure 2.

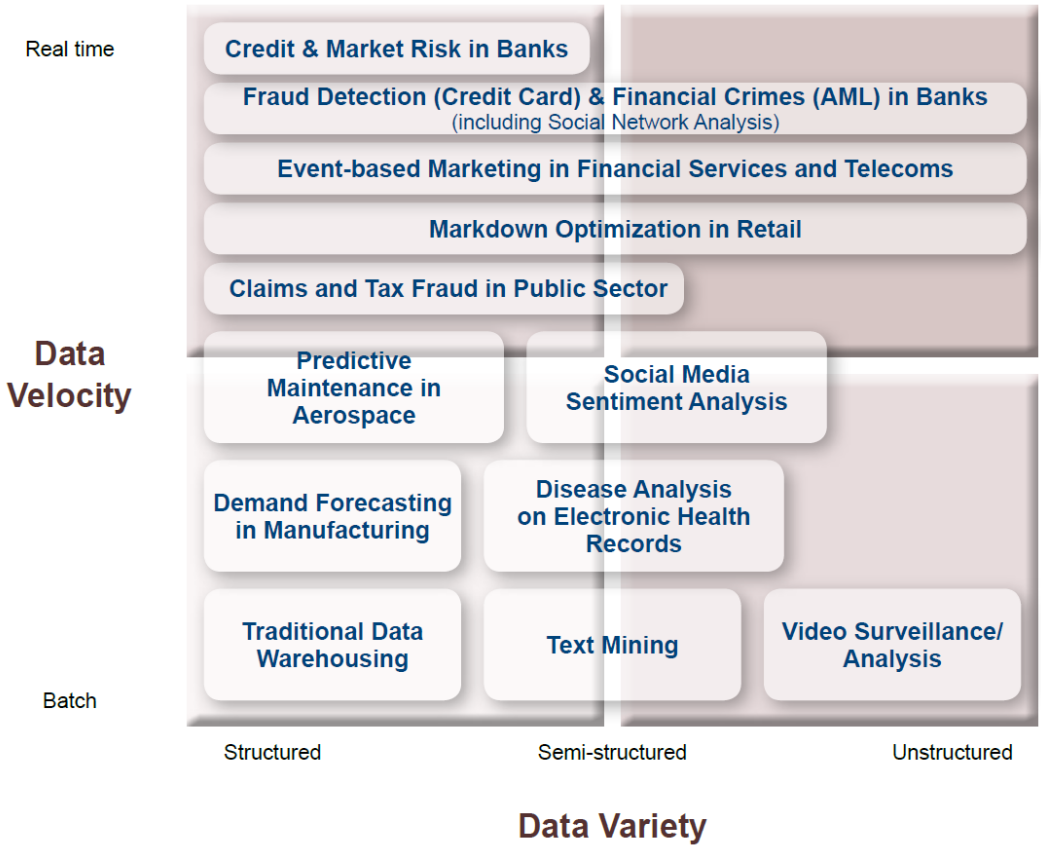


Figure 2: Use Cases for Big Data and High Performance Analytics. [2]

It is important to mention that analytics can retrieve useful information from data that may represents insights. With further analysis can be potentially transfer into foresights.

New emerging area of analytics is represented by unstructured data (text) with broad use of social media. **Text analytics** identifies and extracts the relevant information and interprets, mines and structures it to reveal patterns, sentiments and relationships within and amongst documents [3].

- **Automated content categorization** makes information searches far faster and more effective than manual or retrospective tagging methods.
- **Ontology management** links text repositories together, enforcing data quality with consistent and systematically defined relationships.
- **Sentiment analysis** automatically locates and identifies sentiment expressed in online materials, such as social networking sites, comments and blogs on the Internet, as well as from internal electronic documents.
- **Text mining** provides powerful ways to explore unstructured data collections and discover previously unknown concepts and patterns.

2.5 Business Intelligence

BI and OLAP typically specialise in querying, reporting, and analysing historical data to understand and compare results to date or for specific time periods in the past. Organizations can use BI and OLAP calculations to project a view of what the numbers say is likely to occur in the future. However, advanced analytics can provide an even deeper understanding of why and a scientifically based, predictive view of the future. Advanced analytics provide users with the ability to explore many variables to refine insight. To provide this deeper level of understanding, advanced analytics often need to explore raw, detailed data rather than smaller samples and aggregations, which are customarily used for BI and OLAP.

BI systems offers user interactions through dashboard interfaces that integrate data access and visualizations such as charts and graphs with alerts, indicators, and other changes trackers. Whereas traditional BI reports sometimes provides only static and limited views of historical performance. Modern BI systems can refresh data in dashboards more frequently, allowing users to track metrics that can alert spikes, dips, or other deviations from expected norms in something closer to real time.

What BI systems lack is both the deeper, more exploratory perspective that advanced analytics can provide, and the insights driven by predictive and other analytic models. By interacting with dashboard portals, BI users can consume advanced analytics through visualizations, and use data discovery capabilities to gain a “why” understanding of what the BI performance metrics are showing. Organizations can go further and make advanced analytics operations themselves the drivers, and implement BI dashboards and metrics to provide views into the results of the analytic operations. Examples include analytics that provide insight into customer satisfaction, success in fraud prevention, and so on. [7, 14]

2.5.1 BI Vocabulary

This section explains terms used in Business Intelligence. In the rest of document, the references to these terms are used.

Extract, Transform and Load (ETL)

Data integration technology is generally used to extract transactional data from internal and external source applications to build the data warehouse. This overall process and the steps in it are referred to as ETL for extract, transform and load. The data is extracted from its source application or repository, transformed to the format needed for the data warehouse, and then loaded into the data warehouse. Data integration technology works hand-in-hand with technologies like Enterprise Information Integration (EII), database replication, Web Services, and Enterprise Application Integration (EAI) to bridge proprietary and incompatible data formats and application protocols. [14]

Data Warehouses and Data Marts

A data warehouse or data mart stores tactical or historical information in a relational database and allows the user to extract and assemble specific data elements from a complete dataset to perform a variety of analyses. The data warehouse can be architected according to schema (star, snowflake, etc), data composition (values and attributes) and dimension levels, and descriptors. Data marts enable additional segmentation within a broader data warehouse environment. [14]

Query and Reporting Tools

Most BI systems allow users to perform historical, "slice-and-dice" analysis against information stored in a relational database. This type of analysis answers "what?" and "when?" inquiries. A typical query might be, "What was the total revenue for the eastern region in the third quarter?" Often, users take advantage of pre-built queries and reports. [14]

On-Line Analytical Processing (OLAP)

OLAP analytical engines and data mining tools allow users to perform predictive, multidimensional analysis, also known as "drill-down" analysis. These tools can be used for forecasting, customer profiling, trend analysis and even fraud detection. They answer "what if" and "why?" questions, such as, "What would be the effect on the eastern region of a 15 percent increase in the price of the product?" [14]

2.6 Visualisation

Visualisation is important in this topic because of delivering information for business users. Visualisation helps to understand and gain insights from data [5]. Especially it becomes relevant for finding relationships among thousands of variables.

Visualisation is a problem to select the best visual for data with respect to size, cardinality, type of information, targeted audience, targeted processes. Various graphs, charts, diagrams can be used for visualisation. Simple visuals are for instance line graph, bar chart, scatter plot, bubble plot, pie chart, etc. Each of them has pros and cons depending on data attributes: amount of data points, number of categories, measures and dimensions and of course an information that should be conveyed. [5]

2.7 Reasoning for HPA

This chapter starts with motivation of problem with unlimited growth of data, explaining the knowledge and execution gaps. Business cases describes the real use case and needs of data analysis in business domain. Business Intelligence supports data analysis and reporting, following the business drivers to extract insights from data and eventually transform them into foresights.

Requirements and demand have changed, pre-defined dataset are not enough to sustain the growth of data in different formats (structured, semi-structured, unstructured, geographical, visual) and semantics (transactional, descriptive, historical). Processing of data in batches is complement with proactive and dynamic processing (real-time where appropriate). Data analysis has started with predominantly historic concerns (exploration of primary data) and now it is challenged with predictive, forecasting, optimisation approaches.

Big Data has been formalised for a decade (see next Chapter 3). It changes the attitude towards analytics by creating new problems and limitations of traditional analytical methods. High Performance Analytics is emerging topic that thrives innovation for solving the problems of Big Data. It represents methods how the Big Data can be handled in practise (see Chapter 4).

3 Big Data Phenomenon

As it has been mentioned before, the main raw material in this topic are data, Big Data. In this section, the Big Data Phenomenon is approached from its starting points and causalities. As the growth and volume of data appeared as a remarkable problem for capturing, handling and processing, it has been reactively described by many authors.

Later in this section, a definition of Big Data is provided (as it was initially described by Gartners). Additionally, Big Data can be characterised by four dimensions (4V). Dimensions are used to materialise notion of the Big Data.

Once the phenomenon is defined, discussion about its influences and impacts may start. How does Big Data influence the information infrastructure and technologies? What are the impacts on data storing and data processing? How are the processes and architecture of information and analytics systems affected? Where are the trade-offs and dependencies between operational problems and risks on one side, and innovative possibilities and opportunities on another?

Finally, the missing part to the phenomenon is the way how to handle Big Data. That is evolving questions: Are there already existing solutions to solve the Big Data Phenomenon? What are the architectures of these solutions? Which technologies are used, or which has been designed for this purpose?

3.1 Starting points

How did everything start? Early in 80's and 90's, the first information systems (IS) started exploiting in enterprises and organizations across various industries. Information systems slowly generated more and more data. Enough data that the sense to examine, search and analyse them for information become genuine. Information that could unveiled trends, dependencies, causalities and hidden patterns.

Generated data remained untouched until the evolution of advanced information systems reached the maturity level to be able to effectively process data with analytic methods. Availability of memory (especially with direct access, random access memories) and computation power had been rapidly evolving.

Considering both starting points, enough data and enough data processing capacity, what are causalities of this phenomenon?

3.2 Causalities

Let's first look at the causalities in data processing perspective. Many vendors came up with solutions, mostly monolithic, that were able to setup infrastructure for collecting and managing data. Main idea was to pump everything (generated data) in one central storage. So called Data Warehousing (composed of storages: Data Warehouse, Data Marts) became a core paradigm as a main method and technology for extracting data. Many big sized enterprises and companies adopted it hoping that it would bring them insights and overviews of data. Data Warehouse should have contained all information needed, the one version of the truth, and make it available for business analysis and decision making.

Understating the goal

However, what if business data changed the form, format, syntax, semantics, what would happen then with Data Warehouse solution to which all data had been pumped up? Are there

any needs for its extension, modification, recreation from scratch? Projects with building Data Warehouse have become sprints on long-term runs.

But what if the data continue growing and accumulating, how would the analytical methods applied on data with unremitting growth perform? Furthermore, business requirements on results of analysis have been shifting along with the alterations in business.

Initially, consideration of goal was in extracting, transforming and loading data into central point and consequently utilizing analytical processing to unveil information for business users. Later, the goal was to process all data. IT managers were contented. What about business managers? Everything in this process seemed to be cool – infrastructure, technologies, processes, etc. Except that it was not meant to build an information fortress. It the end, business users seek for concrete information that can be converted into added value for business, competitive advantage and profitability. Having accurate and valuable information has been identified as the crucial target.

Information boom

Phenomenon is powered by multiple aspects related to data and/or information. Let us discuss further causalities, such as sources, structures, growth and relevance of data and/or information.

IT systems have become broadly and enormously incorporated in our daily life. All systems, devices, software that generate any data can be considered as **Data Sources.** Internet as unlimited data source (text, multimedia content: pictures, video, documents), mobile technologies (phone calls registry, messaging, location sampling), systems in any industry, transportation, medicine, services, government, trading (trades, prices moves, bids and offers, clearing data) generate data. Furthermore, some data, let us call them primary, generate another secondary data. Information is shared nowadays faster than human can even realise. Supply of data has exceeded their demand.

Apart from various sources, data are generated and offered in variety of different **Data Structures**. Just imagine melange of data in different forms and formats; with different syntax and semantics; having different life cycles and quality. Thinking about unique and monolithic solution for supporting all kinds of data is almost impossible.

Speaking of **Data Growth**, is there any limit how far the data might breed? In 2012, the 2,5 exabytes of data were emitted every day in 2012 according to estimations by IBM [12], roughly 1 zettabyte of data. This rate almost doubled in last 2 years, mostly affected by massive use of social networking (400 million tweets a day on average, 500 million daily active users on Facebook), spread mobile technologies (4 billion of mobile phone in use). And it is not only a volume, the complexity of data grow as well.

Among all those data, the importance of extracting the information has become substantial, e.g. for organisations to take a competitive advantage, in medicine to explore and predict diseases, etc. **Data Relevance** has become a challenge for anyone who wants to create the most beneficial and profitable information that could be absorbed and effectively utilized by users. These causalities stress the necessity of definition of Big Data.

3.3 Definition

What is Big Data? Term Big Data has been described by many authors (first by Garnter in 2001, later updated in 2012). Initial definition of Big Data is following:

"Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight discovery, decision making and process optimisation."

Garnter

Definition can be interpreted as a relative term describing the circumstances for existence of Big Data. It indicates dimensions of data: volume, velocity and variety that will be described later as "3V". It refers to data as information assets. Second part of definition demands for enhanced insight and accurate timely decision making. This implies a relative absence in supply of storage and computation capacity.

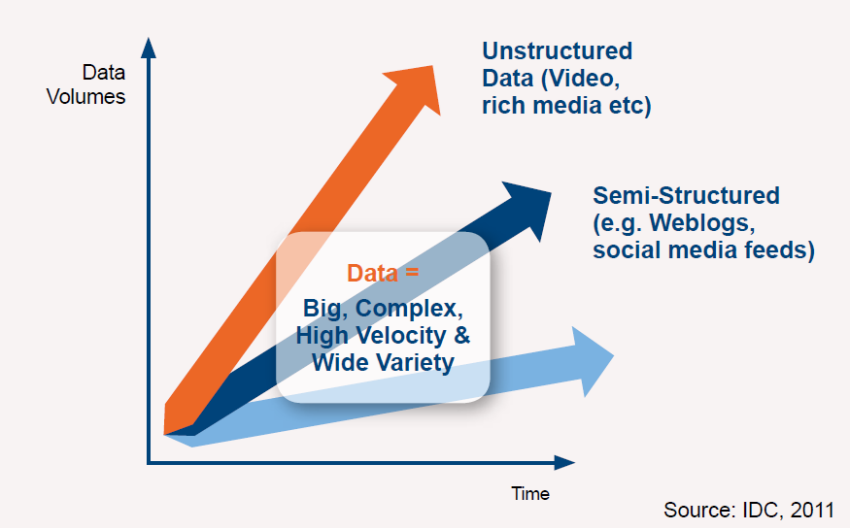


Figure 3: Defining Big Data. [2]

3.4 Dimensions

Definition of Big Data indicates volume, velocity and variety of information assets. Those terms are characteristics called "3V"dimensions. All dimensions are continually moving targets, therefore Big Data can grow in any, if not all, dimensions.

Additionally, one research has extended set of dimensions by Validity Dimension to make sure that content of information is adequate to meet the quality expectations. Another research has introduced Complexity Dimension to underpin the difficulties in processing and extracting the information and demands in available processing capacity and analytical methods.

3.4.1 Volume

The **Volume** Dimension represents physical volume of data, moreover the extremely growing volume. Corresponding growth of data is faster than growth of available storage capacity.

3.4.2 Variety

The **Variety** Dimension represents information expansion resulting into multiple data types (textual, numeric, etc.), formats, structures (structured, semi-structured, unstructured), encoding, syntax, semantics, etc. Up to 85 percent of an organization's data is unstructured – not numeric – but it still must be folded into quantitative analysis and decision making. Text, video, audio and other unstructured data require different architecture and technologies for analysis.

Variability can be comprised with this dimension. Variability should represent inconsistency in data flows. Information can flow from different multiple locations, they can vary within time (with daily, seasonal and event-triggered peak loads that can be challenging to manage). Variability also refers to patterns and relations among information melange and those may be a subject to change.

3.4.3 Velocity

Along with accelerated growth of data, this dimension represents **velocity** of data change. Some data can be changed dynamically after they were created, some can be nearly obsolete in a few seconds. In order to extract information or change information the analysis on these data has to be performed immediately. This requires available allocation of capacity for near-real time processing by analytics methods. Due to dynamics and aging of information this dimension is also known as **Volatility**.

"Initiatives such as the use of RFID tags and smart metering are driving an ever greater need to deal with the torrent of data in near-real time. This, coupled with the need and drive to be more agile and deliver insight quicker, is putting tremendous pressure on organizations to build the necessary infrastructure and skill base to react quickly enough" [Thornton May].

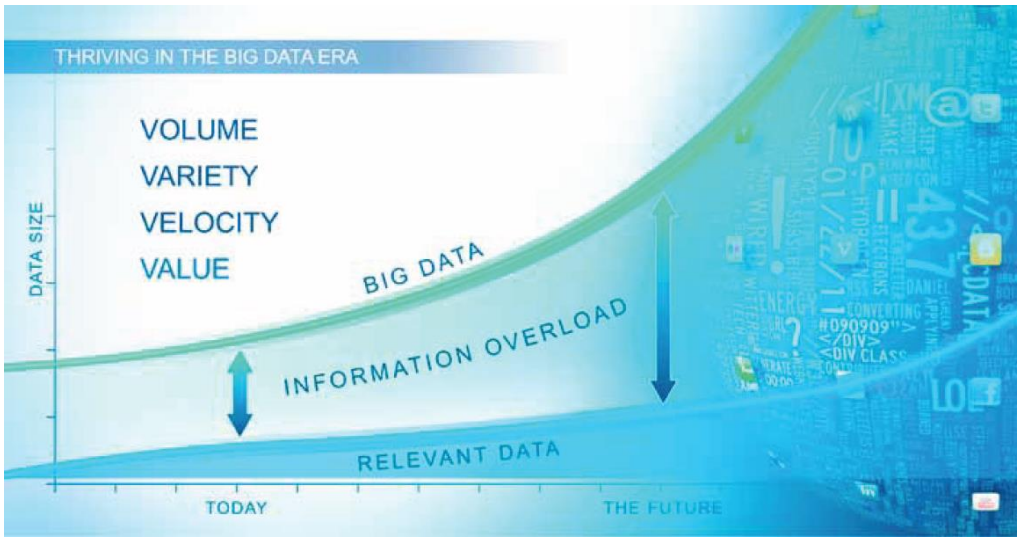


Figure 4: Determining relevant value from massive amounts of data. [4]

3.4.4 Value

Data flow has accelerated in all previously mentioned "3V" dimensions. Obviously, value of data are unequally balanced. Data about customers and orders has different value that data generated by GPS devices. Data can be transactional, descriptive, or historical. Each of them has different value, hides different information.

Value dimension tries to characterise Big Data from perspective which information are useful and meaningful. How is usefulness of information determined? Validity and/or relevance of information is missing with regards to business needs. Hence this dimension can be also called Validity.

Aspects of Value and Validity can be defined as:

a) value of information – how important or crucial the value is for business (not) to possess the information (essential to possess, nice to possess, …)

b) availability of information – how important or crucial the access to information is (not) for business (real time, near-real time, on request, …)
c) stability and aging of information – how important the change of value in time is (not) for business, e.g. changes can be interpreted into patterns, trends, discrepancies

3.4.5 Complexity

The Complexity Dimension was added later to extend the dimensional model of Big Data. In terms of difficulties in dealing with data, complexity, which is implied from other 4V dimensions, increases with the expanding universe of data sources and are compounded by the need to link, match and transform data across business entities and systems. Encouraging organizations for needs to understand relationships, such as complex hierarchies and data linkages, among all data.

Dimension	Brief description
Volume	▪ From terabytes to petabytes and up (zettabytes) ▪ Extremely fast growth
Variety/Variability	▪ Expanding information in Types, Sources, Formats, Structures, Encoding, Syntax, Semantics ▪ Inconsistent data flows, variable in Locations, Times (Peaks)
Velocity/Volatility	▪ Accelerated data flow in all directions ▪ Dynamics, Aging of information
Value/Validity	▪ Useful, Meaningful content of data flow ▪ Quality, Importance, Relevance of information
Complexity	▪ Need to correlate and share data across entities

Table 1: Overview of Dimensions characterising Big Data.

3V is a term referring to volume, variety and velocity dimensions. 4V is a term referring to 3V and including validity dimension.

3.5 Influence and Impact

Explosion of data conveys certain consequences. This section tries to address influence of Big Data on information infrastructure and technologies, impact on data storing and processing methods, influence on processes and architectures. In the end, there is a discussion initiated about trade-offs and dependencies between operational problems and risks on one side, and innovative possibilities and opportunities on another.

3.5.1 Data Flood

Let us try to create analogy between data explosion and floods. Naturally, BI solutions are designed to capture, process data and analyse data. Frequent flooding in landscape initiated construction of dam across a stream, river, or waterway for the purpose of confining and controlling the flow of water. Certainly every constructed dam has own limit of reservoir. It turned out that this solution was not ideal, and problem of flooding persisted.

Later on, supplementary solutions were added into the flood control systems, such as levees, polders, reservoirs, and weirs, so the hydrology control systems could serve for multiple purposes, for instance water supply, hydropower, irrigation, recreation, sedimentation control or flood control. Hence, the consequences of natural hazard have been turned out into gains. Speaking back of data, the volume, variety, velocity of data are not issues rather than it is the data processing.

3.5.2 (Data) Infrastructure and Technologies

The data warehouses and data marts, as part of traditional BI infrastructure that were designed to store "ultimate version of truth" and supposed to support data analysis, are complemented by other means implemented on advanced emerging technologies, e.g. analytical sandboxes, Hadoop, streaming and complex event processing (that are described later in this document).

The volume and variety of Big Data (mostly unstructured data) influenced the evolution of new data storages with alternative storing methods and data accesses (Hadoop, NoSql, Casandra, etc). Data storages has started being optimised for concrete data types as well as the operations on data.

The **velocity** of Big Data motivates the speed of data processing to reach the (near) real-time analytics results. Edge technologies incorporate advanced computing techniques (grid computing, massively parallel processing, rule-based systems), tweak the architectures (in-memory database, in-database processing), thrive manufacturing new appliances (pre-configured solution with tight integration of hardware-software) and/or possible combinations of all above.

3.5.3 Primary/Secondary Data and Processing

There are multiple strategies when processing data. Secondary data are generated out of data analysis from primary data (usually generated from OLTP sources). **Priority Processing** (on-line processing) of primary data meets the requirements when primary data should be processed in real-time (sometimes 5 minutes old data are not valid anymore) or near-real time, when value of secondary data is critical for business and it is intensively used and shared having many subscribers.

On the other hand, **On-demand Processing** (off-line processing) of primary data meets the requirements for campaign, monthly or quarterly processing with sporadic on-demand usage and sharing among a few subscribers.

The influence of Big Data is on processing of primary data. Data with highest relevance for business can be handled with priority on-line stream processing while the rest can be stored temporally in lower cost depended platforms. Together with purpose of secondary data processing (purpose of input for analytical applications and systems) the data will flow into adequate specialised storage systems (unstructured or semi-structured storage for textual data).

3.5.4 Visualisation and Presentation

Big Data brings new challenges to **visualisation**. Visualisation of Big Data needs to take into account large volumes (size of data), different varieties (structured, semi-structured, unstructured) and varying velocities (speed of data flows and frequency of changes) as well as cardinality of columns (number of unique values per column).

Large **Volume** challenges a collapsed and condensed view on results in intuitive and interactive way, availability of results through various channels (e.g. mobile device), and ability to explore data in easily and in near-real time. Examples of solutions how to address the volume of data can be a binning technique (grouping data on all axes in plot) or a box plot (displays more statistics in one plot: minimum, lower quartile, median, upper quartile, maximum). [5]

Different **Variety** brings another challenges. Whereas visualising structured data (measures with dimensions) is reasonably simple the semi-structured and unstructured requires new visualisation approaches. Example of solutions can be a word cloud that displays frequencies of used words or a network diagram that shows relationship among entities (person's followers, friends' network).

A key challenge about varying **Velocity** is to access, process and display quickly how data are flowing into system (or company) and how data change. An example of solution can be a correlation matrix. The correlation matrix combines data and fast response times to quickly identify which variables among the thousands or millions are related, eventually it shows the strength of relationship between variables. [5]

A **filtering** can be consider as another challenge with regards to presentation of Big Data. The task is to seek an effective solution which would quickly filter massive amount of data. Using histogram can lead to better understanding of the composition of data and filter data.

3.5.5 Processes and Architecture

Idea of Big Data Phenomenon forces to re-think the processes and architectures of Business Intelligence solutions. Data storing methods has become a part of infrastructure (see Section 3.5.2).

Traditionally, **Data-Driven approach** in constructing BI architecture involving data (extract transform load), add analytical methods and tools and then derive information for targeted users. Paradigm of this approach was to build up a data storage without complete set of requirements for the analytical output (secondary processing). End user that relies on the results of analytical calculations – information – ended up with having less flexibility and ability for a change requests of analytical output (new requirements for might reconsider entirely the processes of data handling in this architecture). Only availability of data (e.g. in data-warehouse) that could produce information is not sufficient.

Since business users have become aware with evolution and power of technology, they are able to request concrete information they want to get as the results (of secondary processing). **Information-Driven approach** in constructing BI architecture involved an information design process that define the architecture. Information design process is driven by **business requirements** that denote required and expected output – information that the business users seek for. Analytical methods are selected accordingly to transform data into required output. Supplementary, requirements should define also the structure of output results and access to these results, which implies the requirements on data storage management system. This approach is also known as **Analytics-Driven approach** with simple idea behind: storing data based on their further processing.

3.5.6 Problems, Risks vs. Possibilities and Opportunities

With respect of the influences and impacts of Big Data on infrastructure, technology, data processing, processes and architecture, the problems and risks pop up.

Firstly, the idea of monolithic solution of BI is no longer a solution for Big Data, new types of data storage evolve and enrich the infrastructure and architecture of solution for Big Data (see Section 3.5.2). This means that consistency of data ("ultimate version of truth" with monolithic solution) is traded off with flexibility and ability to process different data with higher speed.

Secondly, with Information-Driven approach of processes and architecture (see Section 3.5.5), when design is proposed by output information requirements, some data are missing (filtered out) and thus information can be lost. Whereas with Data-Driven approach, the process is designed to store the most relevant data, ideally all, which is costly (capacity requirements for ETL and storing).

Thirdly, with monolithic solution of BI the organisation is usually dependent on one vendor, on its proprietary technology, which obviously leads in higher costs, low efficiency, and low motivation for innovations. Since the monolithic solution is challenging with emerging technologies (distributed data-warehouse) the organisations may opt for another, combined solutions from multiple vendors.

All problems or trade-offs mentioned above create opportunities for more elegant and intelligent solutions (Hybrid Architecture, Upstream Analytics, see Section 3.6.1). Growing data in all 3V dimensions creates possibilities for enhancement of information architecture and technology, for changing philosophy of data handling and capturing, for perception of information by human.

To sum up, the issue is not a growing volume of data rather than processing of all available and relevant data. The data overload (supply of data breach the information demand) is not an issue rather than producing the information that can be perceived by human, systems or artificial intelligence. Availability of data is not an issue rather than data organisation and management.

3.6 Approaches to handle Big Data

Big Data can be seen as problem, on the other hand as opportunity (see previous section). This section contains approaches of handling the relative big data from perspective of architecture, processes, infrastructure, and technologies.

3.6.1 Approaches

The traditional BI architecture can be considered as a starting point for architectures with its process (including staging area, data-warehouse, data marts, ETL, etc.). Looking for limitations of this architecture, one may find out that it is unlikely to store all data in central (enterprise) data-warehouse and not all data are necessary to be stored. [10]

There have been new architectural approaches evolved: **Hybrid Storage Architecture** (combination of storages for various data types and formats, temporary data storages, data stream processing), **Upstream Intelligence** (analytical and statistical functions are applied early in the process during acquisition of data) that includes also specific **Stream and Event Processing** (based rule-based systems, pattern identification).

As results of this evolution, **Post-modern BI Architecture** represents a complex solution that has been inherited, mostly from traditional Business Intelligence and adds the concept of hybrid storage architecture, upstream intelligence, and stream an event processing. Post-modern BI Architecture consists of distributed data-warehouse, consolidated meta-data layer, coordinated management of data streams (ETL) and collaboration knowledge management.

Hybrid Storage Architecture

Hybrid Storage Architecture represents an idea of combining a traditional data-warehouse (designed for structured data) with various data storage for different types and structures of

data and analysis applications (Hadoop, NoSql). The concept also considers temporary data storages that buffers data, additionally data can be discarded after the certain time-out. Critical information (content and time wise) can be stored on platforms with advanced technologies, directly supporting analytical processing with scalable performance. The less relevant information or not time critical information can use simpler and cheaper technologies. To sum it up, different destination for different data is determined according data priorities and further processing methods.

Upstream Intelligence

Idea of Upstream Intelligence is simple: apply analytical processing and methods (typically included as a last bit of process - downstream) in the initial phase of extracting data from data source - upstream. The aim is to control data that go into data management systems by filtering out irrelevant information, hence data can be evaluated immediately (by using analytical and statistical functions) accordingly to the importance and relevance prior to being stored.

Technically, analytics can be plug into the ETL processes (user defined analytical processes can be deployable in upstream). This allows to immediately analyse quickly aging data (from perspective of business). Therefore Upstream Intelligence can be wired with the processing of data streams and events, which is implemented by Complex Event Processing (CEP) technology. CEP supports continuous intelligence and they are used as intelligent sensors that can be attached to streams with large volume of data and monitor combination of events in (near) real time.

3.6.2 Post-modern BI Architecture

Due to diversity of requirements from business orthogonal BI architectures evolved: a Top-Down and a Bottom-Up architecture. The **Top-Down architecture** stressed out a report-driven or a data-driven approach where a data warehouse model is created first based on the business/reporting requirements. Process of this approach starts with an ETL routine to move data from source system to the data warehouse (DW), and then continues with creating reports and dashboards to query data in DW. This approach mostly satisfies casual users with periodical reporting and monitoring. [7]

Apart from that, organisations pursuits power users to work on ad-hoc analysis or tasks in research and development department. With previous approach power users are left aside to use ad-hoc spreadsheets, separate/local database instances, SQL and data-mining workbenches. With Top-Down approach power users find BI tools inflexible and a data-warehousing structure too limited for their concerns. Opportunity for Bottom-Up architecture approach has appeared.

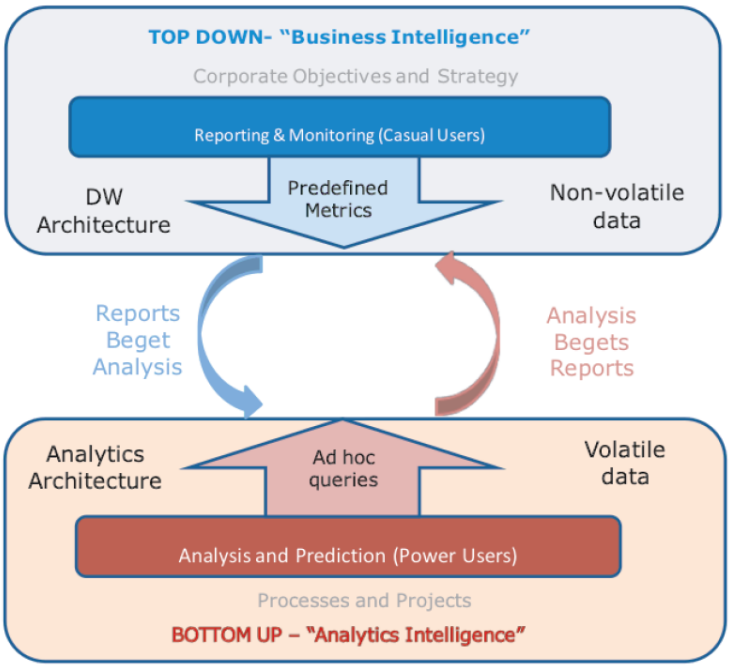


Figure 5: Top Down vs. Bottom Up architecture approach. [6]

The **Bottom-Up** approach suits better for business analysts and data scientists who require the ad-hoc exploration of any data source, both inside and outside corporate boundaries, working closely with business managers to optimise existing processes. [6]

Post-modern BI architecture is a result of expansion of data warehousing architectures, data governance programs and adding advanced analytics in order to balance the dynamic between top-down and bottom-up requirements. This architectural concept is also known as **Hybrid architecture** (depicted in the Figure 6). Big Data and HPA do not change data warehousing or BI architectures. They simply supplement them with new technologies and access methods better tailored to meet the information requirements.

Hybrid architecture can optionally contain following complementary technologies (described more detailed in Section 3.6.3) like:

- **Hadoop clusters** – to support storage for semi-structured data, used in staging area or analytical sandboxes
- **Streaming** and **Complex Event Processing Engines** – to support continuous intelligence, used as intelligent sensors that can be attached to streams with large volume of data and monitor combination of events
- **Analytical Sandbox** – to boost analysis processing, ad-hoc queries, to satisfy short-term analysis needs, used as an access points for other BI systems
- **Non-relational database system** – to store unstructured or raw data, used in analytical sandbox, or staging area
- **Data hub** – to feed other systems and applications rather than to host reporting or analysis applications directly, data-warehouse used as a hub

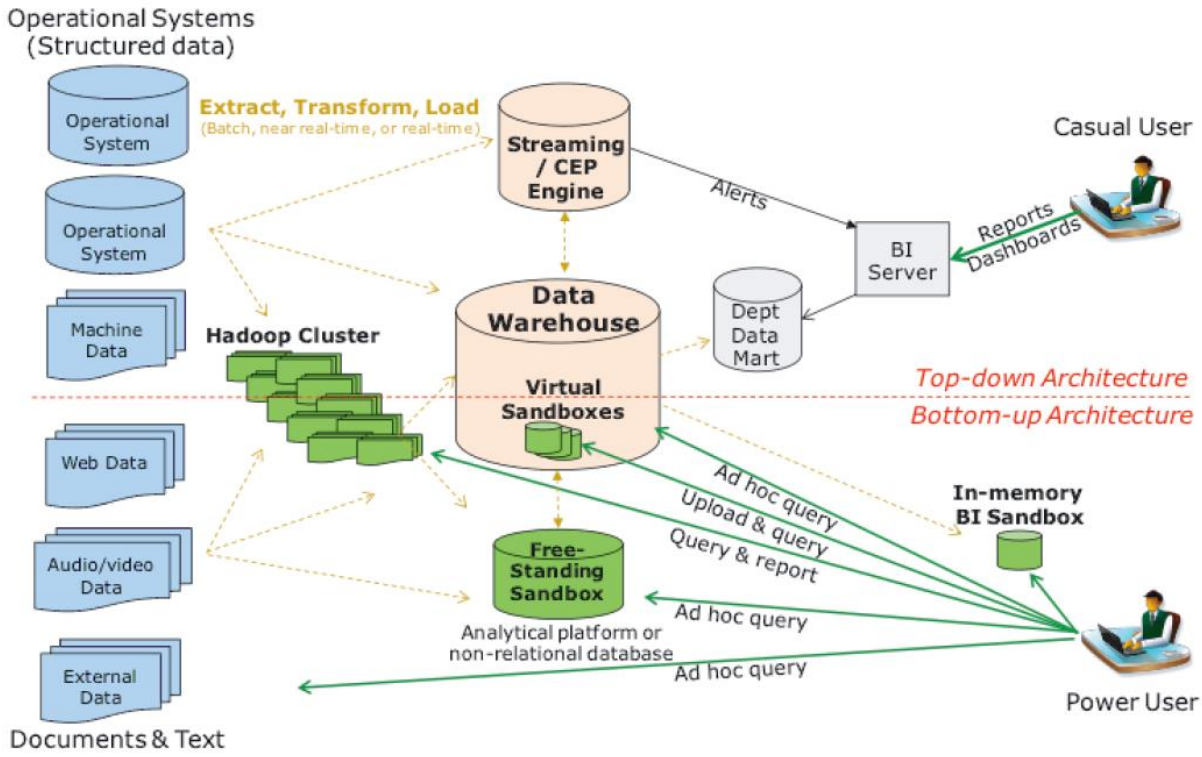


Figure 6: Post-modern BI architecture. [6]

3.6.3 Architectural Features

Speaking about architecture for High Performance Analytics, various concerns pop up. What king of storage is used for different types of data? How are these storages integrated into the information infrastructure? What is the role of data-warehouse, how is it primary used? How are the advanced analytics plugged into architecture in order to be able to access data and perform analysis processing? How are the data streams handled?

All of those are concerns of HPA architecture that needs to be addressed by architectural features. There are various features that can be found and utilised in any or both architectural approaches: Hybrid Architecture and Upstream Analytics. In the following text, these features are described in details.

Data Storage and Integration

In Post Modern BI architecture, a data-warehouse is no longer a centrepiece. Increasingly it is accommodated along with another systems that manage structured and unstructured data. However, the main part of data-warehouse architecture keeps running on relational database engine in integration with Hadoop and non-relational database.

The **relational** databases and **RDBMS** (relational database management systems) are designed to store data in structured way (as a collection of tables that are formally described and organized according the relational model). Relational database is mostly used as data-warehouse.

Hadoop is an open source framework for building data-intense applications. It runs on special distributed file system. Since it is a file-based, data model is not needed, it is “schema-less”. Hadoop can be used to store and process any kind of data (transactional and structured, semi-structured, unstructured video or audio). It can be used within BI architecture as: **a)** online archive, as **b)** staging area with and support ETL process by parsing, integrating and

aggregating large volumes of data and shipping them to the data warehouse, as **c)** as an analytical engine for running analytical computations against large volumes of data (queries can be implemented in Java and according MapReduce), as **d)** scheduled reporting engine that run against raw data instead of summarized and aggregated data (report querying can use Hive or Hbase and return data to BI tools). Disadvantage of Hadoop is that it is batch-oriented and not conducive to iterative querying.

Non-relational databases are used to store structured and unstructured data within a single index and to give user unified access to query any type of data providing search-like characteristics. Typically, entities are extracted from documents, files and other databases using natural language processing techniques, and then are indexed as key-value pairs for quick retrieval using a document-centric query language such as XQuery. Non-relational databases support iterative queries so they can be used to monitor trends. These non-relational systems complements Hadoop, an enterprise data warehouse or both, because they are capable to combine and correlate data coming from both.

Back to integration of diverse data storage, Hadoop is suitable fast data loading for staging area. Afterwards, it can behave as online archive for raw data, however the summary of that data can be loaded into data-warehouse. Non-relational database boosts mostly analytics complementary with Hadoop. Data-warehouse in this integration holds transactional and structured numeric data along with maintenance of constituent metadata.

Data Federation

Another approach of integrating data is described in [7]. So called **Data Federation** is a data management strategy to support analytics and to provide real-time access to the data that are physically located in diverse sources. Data Federation allows to virtually join data together, even though data of heterogeneous types (including non-relational data such as legacy file systems, spreadsheets, XML streams, search results, etc.), and all promoted as middleware or façade layer. Thus, data movement and data redundancy are minimised, because data are not moved around or transformed to another separate storage.

Data Federation is an alternative to ETL batch processes. It provides a “loosely coupled” approach to data integration in circumstances where it would be too costly, slow, or constraining to try to centralize the data—especially for near-real-time access. Also query plans must be considered with approach in terms of query optimisation and ensuring that only requested data in requested detail are transferred over the network. Eventually, evaluation time of query and latency can be improved by introducing a cache for recent and/or frequent queries.

Data Hubs

Typically, a data-warehouse (DW) manages structured data from the business operational systems (OLTP), while Hadoop and non-relational databases primarily manage semi-structured (e.g. log files) and unstructured data. The DW is increasingly running on traditional rational databases and it is used as a **hub** to feed other systems and applications rather than to host reporting or analysis applications directly. For instance in this mode analytical queries can be applied against virtual cubes that run in memory or reside in analytical sandboxes. [6]

ELT vs. ETL

An **ETL process**, extract-transform-load, is commonly used in data warehousing to gather data from multiple data sources (tables, streams) and prepare it for a user of BI and OLAP.

ETL processes are typically executed during the off-peak hours, because they require a lot of computing time and power. And data are ready to be accessed when the ETL processes are completed. This downside arises a new idea: to shift some of the transformation workload away. [6, 7]

An **ELT process**, extract-load-transform, pushes the transformation steps into the database engine (it can be the same platform as data-warehouse) where small atomic tasks are converted into SQL statements and procedures. Data is extracted and loaded into a raw format. Afterwards data is transformed and moved into data model that is accessible by target users. While ETL is usually design and deployed as a separate layer of information system, ELT benefits from the power of database engine. Therefore ELT can be scaled out in a sense that most of the steps covering data integration (soring, merging, summarizations, profiling, etc.) are run as ELT processes on database engine. [7]

The ELT process introduces the concept of transformed data on demand by responding to analytical operation requests. ELT is complementary to ETL, the ETL supports to handle regularly scheduled transformations, whereas the ELT is oriented on the real-time or on demand requests.

Analytical Sandboxes

DW, in keeping with its role as a hub, distributes data to analytical sandboxes that are designed to be accessed and examined by ad hoc analysis by business analysts and data scientists. [6] This architectural feature is complementary approach for data marts. There are four types.

Virtual Sandbox – it is a partition (or set of tables) inside the DW dedicated to individual analysts Instead of creating separate data mart, external data can be uploaded into the partition and combined with data from DW that are pushed (by ETL) or pulled there (by queries). With allocated computational resource it creates autonomous analytical unit.

Free-standing Sandbox – is a separate system from DW with own computation resources, storage and database to support complex queries. In some cases, it runs queries against a replica of data stored in DW. In other cases, it runs own non-relational database containing a set of data that does not fit into DW because of scape or data structure issue.

In-memory Sandbox – maintains a local data store in memory to support interactive reports and queries. Data are pulled from any source and quickly linked. Super-fast queries run against these data held in memory supporting fast visual interactions. Existing BI tools are QlickView and PowerPivot.

Hadoop – as a set of technologies, can be considered as an analytical sandbox too, due to its environment supporting complex queries and calculations against raw and atomic data (in contrast with summarised and transformed data stored in DW).

Streaming and CEP Engine

Another architectural feature particularly to support continuous intelligence is Complex Event Processing (CEP) and Streaming Engines.

The **CEP Engine** is designed to ingest large volumes of discrete events in real-time, calculate and correlate those events on patterns and relationships, enrich them with historical data if needed, and apply business logic, rules or analytical models that trigger notifications or alters on specific actions or anomalies (when conditions of rules are met).

The **Streaming Engine** is similar to CEP engine but is designed to handle enormous volumes of a single discrete event type. This engine typically ingests an order of magnitude more events per second than CEP, however only from a single source.

Both engines are part of rules-driven systems. Rules-driven systems behave as intelligent sensors that can be attached to streams of transactional data and watch for meaningful combination of events or trends. The CEP systems are sophisticated notification (rules-driven) systems designed to monitor real-time events. Internal design for analytics that process real-time data and events is depicted in Figure 7. [6, 7]

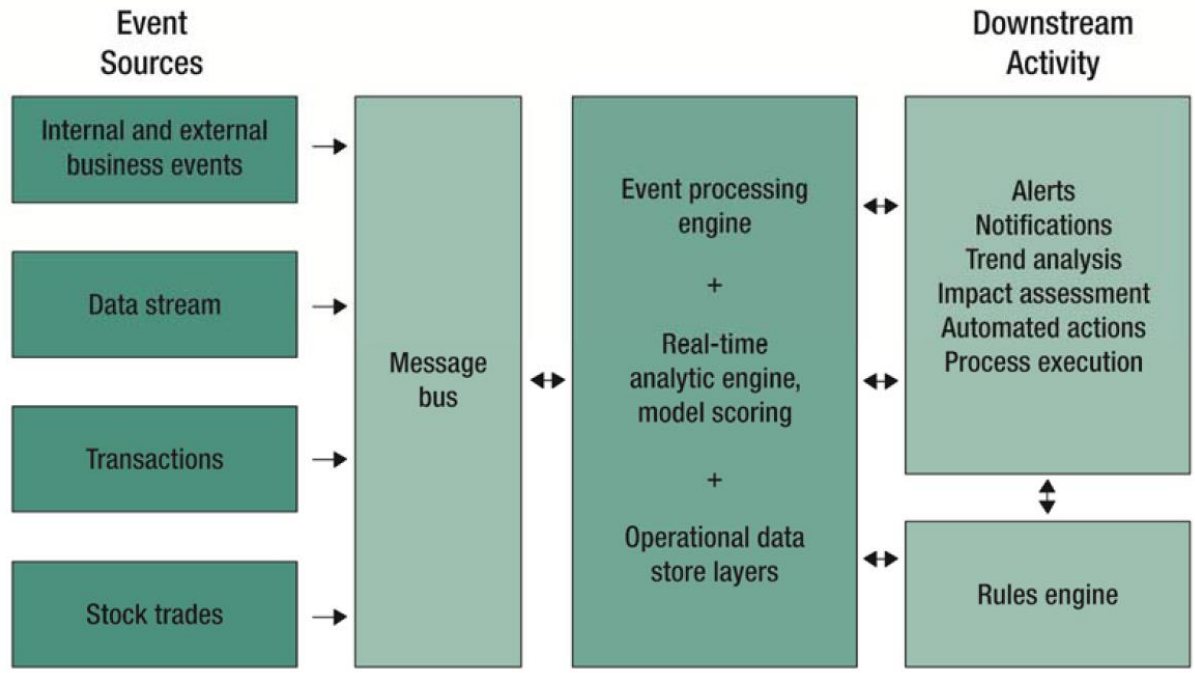


Figure 7: Real-time data and events processing for analytics. [7]

CEP Engine are perfect for detecting fraud in stream of transactions per second, for supporting a logistics optimisation or a traffic regulation in streams from transportation operations and sensors, for regulating the process on a production line in factory, for reacting on market move in streams from stock price tickers, etc.

3.6.4 Technologies

New technologies has hit the market to focus on mainstream of Big Data. In this section the overview of technologies is given with a brief context explanation. A separate subsection is dedicated to specific family of architectures and technologies that address how advanced analytic techniques operate on big data – concept of HPA.

Big Data Technologies

In the following table, the overview of technologies that has been developed recently to deal with Big Data Phenomenon is given. [2]

Technology	Context
Big Table	Proprietary distributed database system built on the Google File System. Inspiration for HBase.
Cassandra	An open source (free) database management system designed to handle huge amounts of data on a distributed system. This system was originally developed at Facebook and is now managed as a project of the Apache Software foundation.
Data Warehouse & Analytical Appliance	Consists of an integrated set of servers, storage, operating system(s), database, business intelligence, data mining and other software specifically pre-installed and pre-optimised for data warehousing.
Distributed System	Multiple computers, communicating through a network, used to solve a common computational problem. The problem is divided into multiple tasks, each of which is solved by one or more computers working in parallel. Improved price:performance ratio, higher reliability and more scalability.
Google File System	Proprietary distributed files system developed by Google: part of the inspiration for Hadoop.
Hadoop	An open source (free) software framework for processing huge data sets on certain kinds of problems on a distributed system. Its development was inspired by Google's MapReduce and Google File System. It was originally developed at Yahoo! and now managed as a project of the Apache Software Foundation.
HBase	An open source (free) distributed, non-relational database modelled on Google's Big Table. It was originally developed by Powerset and is now managed as a project by the Apache Software Foundation as part of Hadoop.
MapReduce	A software framework introduced by Google for processing huge data sets on certain kinds of problems on a distributed system. Also implemented in Hadoop.
Non-relational database / Key Value Store	A non-relational database is one that does not store data in tables (rows and columns) – in contrast to a relational database. Key Value Stores allow for the management of schema-less (noSQL) entities.

Table 2: Overview of Big Data technologies.

Concept of HPA

Although, the evolution of advanced analytics has been driven mostly by new structures of data and demands of business, there are methods for intelligent selection, integration, aggregation, which were developed in context with data management, to facilitate problem with Big Data. The concept of HPA tries to combine advanced analytics and those methods with high-performance computing techniques.

The concept of HPA relies on high-performance computing techniques as Parallel Computing, In-Database Analytics, and In-Memory Analytics. The following table gives a brief overview and characteristics of those methods [3]. However, next Section 4 is completely dedicated to present HPA in details.

Parallel Computing	Context
Preconditions	- Split task and data into autonomous segments - Minimal interactions - Exclusive Datasets
Advantages	- Effective Resource Allocation - Task Prioritisation
Benefits	- Speed up Complex Computing on Big Data - Naturally Segregated Computing
Realisation	- Massive Parallel Processing - Grid Computing – Enable to automatically use a centrally managed pool of resources assigned in a grid environment to achieve workload balancing, high availability and parallel processing.

Table 3: Characteristics of Parallel Computing

In-Database Analytics	Context
Preconditions	- Move Computation Closer to Big Data - Database System (DBMS) - Analytical Methods as Function and Procedures of DB
Advantages	- Computing Resource of DBMS - Minimum Data Movements/Loading - Tuning of Analytical Model directly on Dataset
Benefits	- Speed up Computation and Analytical Cycle (Cycle in steps: Stream it, Score it, Store it) - More Accurate Results of Analysis (queries against raw, not summarised or aggregated data)
Usage	- Processing of Big Data that are associated with specific utilisations in Analytics
Realisation	- In Database Analytics – Tasks moved closer to data, running computations inside the database to promote better data governance and faster results. Analytics run inside, avoiding time-consuming data movement and conversion, constrained by allocation of database workload management system.

Table 4: Characteristics of In-Database Analytics

In-Memory Analytics	Context
Preconditions	- Data Loaded to Memory of Analytical Machine or Appliance
Advantages	- Direct Address in Memory - Complicated Computation, Advanced Statistical Methods - No continuous data loading
Benefits	- Complex Computation - Predictive Modelling, Correlation Analysis - Quick Response Time
Usage	- Advanced Data Exploration - Dynamic Recalculations (of Portfolios)
Realisation	- In Memory Analytics – Divides up analytic process into easily manageable pieces. Computations can be distributed in parallel across a dedicated set of blade servers.

Table 5: Characteristics of In-Memory Analytics

3.7 Discussion

Big Data Phenomenon has been described in this section with its causalities, definitions, influences and impacts.

It is important to mention that the paradigm shift from data-driven approach towards information-driven approach may be seen destructively on the traditional idea of central data-warehouse and physically enforced data integrity and consistency. Rather than that, it creates complementary solutions that satisfy requirements for flexibility and adaptability supporting data analysis. Information-Driven Approach is a starting point for Upstream Intelligence.

Big Data shuffles architecture in that way that Post Modern BI Architecture developed in melange of Hybrid Storage Architecture (Analytical Sandboxes, Hadoop, NoSQL, RDBMS), Upstream Intelligence (particularly supported by Complex Event Processing), and traditional data-warehousing. Another remarkable idea, following the Hybrid Storage Architecture, is Data Federation that encourages to locate data on multiple appliances in various structures and formats. However it supports analytics and ensures providing of real-time access to data via virtually unified data access.

Having the BI architecture exhaustively extended (Post Modern BI Architecture), there are yet technologies that thrive for innovative approaches to handle Big Data. Area of high performance analytics is not comprehensively defined and mapped. The evolution of HPA is mostly driven by business demand for having information (results of data processing) immediately and by availability of resources (grid computing, very large memory with direct access).

4 High Performance Analytics

This section explores the world of analytics, high-performance computing and luckily combinations of both. This section describes the drivers, definitions, classifications, technologies of high-performance analytics.

4.1 Drivers and Boundaries

High-performance analytics (HPA) emerges with increasing demands for supporting advanced analytics and shifting paradigms about data management. Advanced analytics often demand larger, detailed data volumes than the smaller pools of aggregated data frequently used for BI and online analytical processing (OLAP). Data availability becomes crucial particularly when analytic models need to be deployed and worked against real-time or at least very timely data and traditional approaches to data integration, data management, and (ETL) processes may not be optimal for advanced analytics.

Decreasing cost of memory - RAM, SSD, hard drives (price per gigabyte), increased addressable space for memory (64bit operating system), increased computational power of hardware, increasing volume of data (Big Data), demand for real-time data processing, demand for complex analytics. All of them can be considered as drivers towards the concept of high-performance analytics.

The aim of HPA is to improve and facilitate an effective allocation of computation and the capacity resources (processing speed, computation complexity, data storage, network traffic).

Analytics

There are four types of analytics. **Descriptive** analytics explore current trends, customer behaviour, relations, trying to focus on output parameters and to answer question “What is happening?” **Diagnostic** analytics explore causes and reasons within the context, trying to focus on input parameters and to answer question “Why is it happening?” **Predictive** analytics try to figure out the future trends and scenarios, addressing the future output parameters and answering question “What will be happening?” **Prescriptive** analytics going further trying to figure out the optimisation of input parameter in order to achieve the expected output parameters.

Predictive and prescriptive analytics rely on intense calculations and thus require higher capacity in terms of computation resource. Nevertheless, all types of analytics aim to process the largest dataset as quickly as possible in order to improve accuracy of results and that requires capacity in terms of data storage.

Advanced Analytics

Advanced Analytics are comprised with number of practices and technologies, including data mining, predictive analytics, natural language processing, and artificial intelligence such as machine learning, decision trees, and neural networks. Advanced Analytics involve statistical, quantitative, or mathematical analysis of data and developing testing, training, scoring and monitoring models. [6]

Advanced Analytics are used to discover why something happens, what happens next, and how to optimise actions to achieve desired results. Advanced Analytics mostly need to explore raw, detailed data rather than small samples and aggregations (designed for BI and OLAP). [6]

Advanced Analytics, sometimes called **Explanatory Analytics,** offer highly complementary technologies to the Business Intelligence. Analytic models (with phases training, deployment, monitoring) are running against data or event streams and requiring the computation power.

High Performance

High Performance is important in phase of a model deployment. Demands on real-time decisions in operations supported by analytics increase massively data movements and replications. High performance methods can be considered massively parallel processing, grid computing, in-database, in-memory, complex event processing, stream processing, etc.

4.2 Definition

Let's try to find now a synergy between both concepts described above. High performance computing thrives to improve the scale and speed of advanced analytics, including upgrading the ability to deliver analytics through current BI and data warehousing systems.

*The concept of **high-performance analytics** is about these high-performance computing techniques specifically with analytics in mind. It is a bit of nuance, but it refers to applying advanced analytics as a core piece of the infrastructure.* [3]

Big data analytics is where advanced analytic techniques operate on big data. [8]

Therefore concept of HPA methods can be delivered as a combination of **architecture** that facilitate HPA (see Section 3.6.1), **technologies** that create infrastructure for application of HPA (see Section 4.4.1). Combination of both creates **analytical platforms** that boost HPA methods (see Section 4.4.2).

4.3 Classification

There is an extensive research done arguing about what exactly belongs to HPA world. Let us split HPA methods and introduce classification by perspectives. Perspective can be considered a) problems and tasks they try to solve, b) dimensions of Big Data they are trying to handle and manage, c) system resources that boost the performance. Classification and perspective brings superior overview into HPA world, especially how to look at HPA. This classification is introduced as a part of research in this work.

4.3.1 By Problem and Tasks

One can consider to split high performance analytics according to the traditional tasks, which try to solve, into the areas of interest as follows:

- Overview – process all or subset of data or aggregations, not modifying data
- Trends – monitor changes only
- Analytical models – examine raw data and large volumes
- Reporting – online/offline reporting, OLAP processing of aggregated data
- Real-time monitoring – examine data streams and events

The analytics can be classified also according to the type of analytics they excel at (Descriptive, Diagnostic, Predictive, or Prescriptive Analytics). As main problems that the high-performance analytics try to address are continuous intelligence (real-time processing of data streams and events, mostly applicable in fraud detection, aircraft or traffic control, factory production line) and iteratively scheduled batch processing of large data (statistical analysis, advanced exploration, ad-hoc analysis).

4.3.2 By Dimensions

From perspective of dimensions of Big Data, the high performance analytics (specifically the analytical platform underneath) can be classified accordingly. Volume of data is suitably addressed, in HPA for certain purposes, by parallel processing (to improve and speed up data processing), upstream intelligence concept (to filter out unnecessary data), storing data based on their further processing (to balance priorities and structures among data), and storage attributes (to minimise I/O operations, to maximise storage capacity with balanced price per megabyte).

Variety of data in HPA can be addressed with range of text analytics (Sentiment Analysis, Text Mining) specifically on unstructured data. The dimension of velocity of data is supported less than the dimension above, although HPA can be integrated together with event processing and data stream processing.

Realisations and technologies of high performance analytical platform from perspective of Big Data dimensions are summarised and addressed by architectures in Section 4.4.2 Analytical Platforms.

4.3.3 By System Resources

From perspective of system resources, the high performance analytics can be classified accordingly. In-Database Analytics are supported with analytical operations in database engine of DBMS (as functions or procedures). In-Memory Analytics are supported with random access memory where the large amount of data can be loaded at once and performed. With availability of processing power, analytics have support of massively parallel processing and grid computing.

Realisation and technologies of high performance analytics from perspective of performance computation are addressed by technologies summarised in Section 4.4.1 High Performance Computing.

4.4 Technologies

There are two major areas of technologies on which the concept of HPA relies: high-performance computing and analytical platforms (a data management system optimised for query processing and analytics).

4.4.1 High Performance Computing

What about putting advanced analytics and data together? Let us try it. Let us try it. This section focuses on technologies that integrate the analytical operations within the storage management systems (In-Memory Processing, In-Database Processing) and parallel data processing (Grid Computing, Pipelining).

In-Memory Processing

High-performance analytics has been exposed to new possibilities in availability of much larger random access memory (RAM). Recently, the cost of memory has continued to fall while the amount of addressable memory has increased (due to the 64 operating systems). **Very large memory** (VLM), up to terabyte, has become available for BI, analytics, data marts and data warehousing systems. With compression techniques VLM allows to pack more data into memory, to make effective use of specialized approaches such as columnar databases, and to bring data closer to the computation operations of analytics.

With traditional approach, BI tools and analytics run queries against data-warehouse that relied on (R)DBMS reading information from disk storage. Obviously, the I/O operations have become performance bottleneck. Approach with using caches has not completely solved it due to the size (paging and swapping memory) and volatility limitations. By introducing VLM and **in-memory system** (processing and database in memory), analytical models can score locally (even with large numbers of variables and low latency) and can be deploy against the near-real time data that are updated continuously with incremental loading.

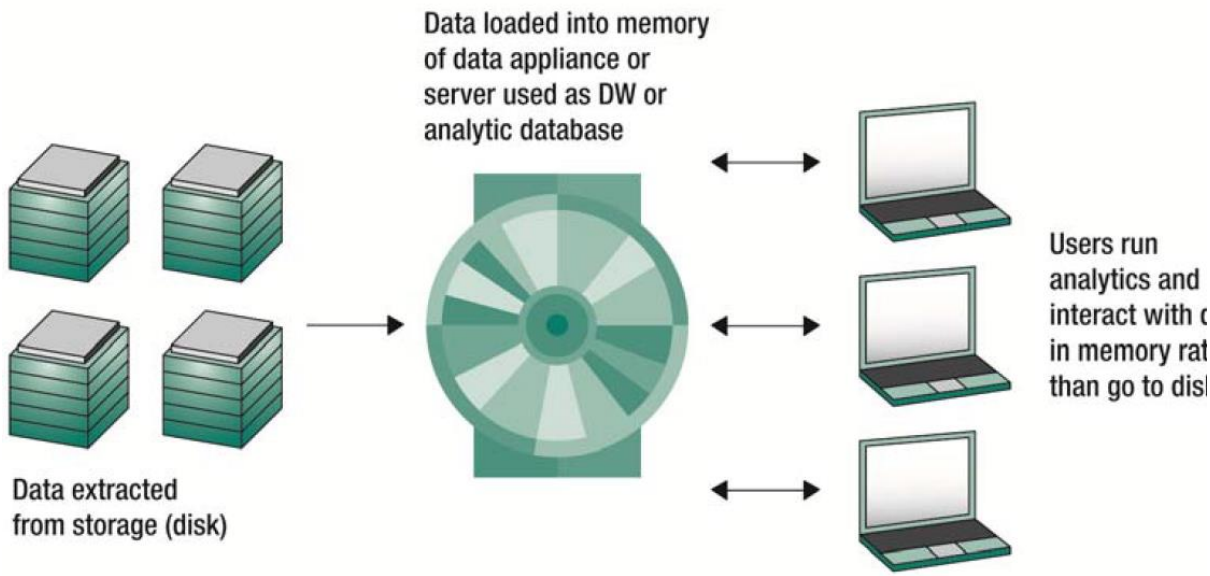


Figure 8: In-memory processing to support advanced analytics. [7]

This technology involves data management facing new challenges. Managing data models and data location is subject for optimisation. Based on an analytical and data model size and complexity the challenges are: how the system manages data stored in rows, columns and tables, how the pointers and vectors to reference data are used in different location in memory, how data are or loaded, how often data are refreshed, how the updates are synchronised.

In-memory analytics are facing challenge in balancing the data compression that can be critical to keep more data in memory. Compression is even more relevant when underlying database system is based on column oriented.

In-memory analytics performance depends on shared memory space management and pipelining. Both are important for analytical process to read and exchange data and execute operations. The benefit coming from properly managed shared memory is that data are easily accessible to multiple processes running in the system and overhead in passing data in traditional architectures (client-server) is reduced. [6, 7]

In-Database Analytics

Organizations increasingly demand for analytics to support (near) real-time decisions which results into massive increase of data movements (each analytical operation is evaluated and results set is being returned) and frequency of analytical model deployment (analytical model is being tuned on-the-fly and has to deployed quickly). [7]

This issue can be addressed by exploiting the SQL database engine or analytical functions of data-warehouse appliances to achieve the level of performance and analytical model management. This approach is called **in-database analytics**. How is this utilised? Analytical model can be translated by data-mining tools into these database engine functions or procedures that execute the modelling steps. The performance gains are achieved when these function can run using MPP (Massive Parallel Processing) database engine. Also improvement in less data duplications and movements is achieved essentially with skipped loading and replication procedures.

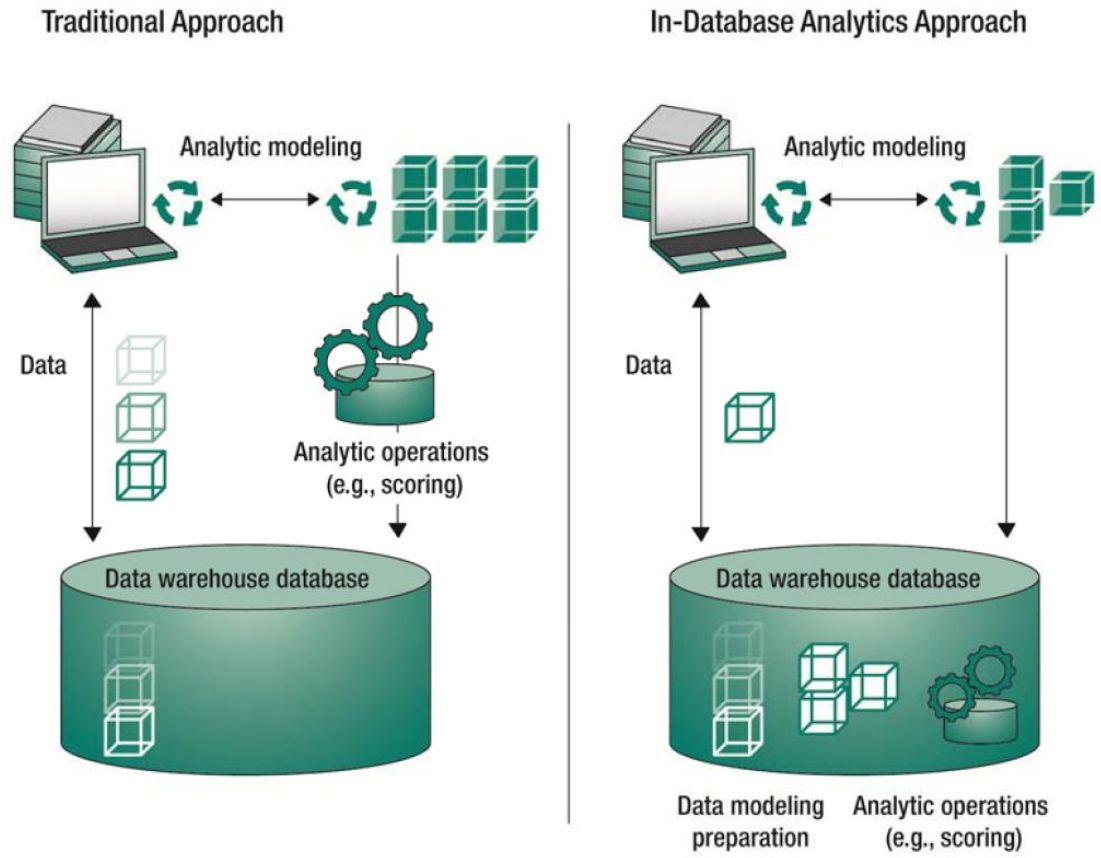


Figure 9: Comparison of traditional approach with in-database analytics approach. [7]

This has two drawbacks. First, the analytical model after modification has to re-develop in round between automated modelling tool and database engine (coding, testing, and validation). Second, variety of model types that can be deployed is restricted by flexibility of conversions from analytical functions into SQL engine functions.

The advantage of this approach is mostly gained when using the database functions for summation and sorting because the yet aggregated or sorted data can be feed into analytical model processing on in-database engine instead of being transferred over network to analytics. Obviously, this approach enables users to access all the data rather than subsets or aggregations.

Parallel Processing

The following techniques and their combination enhance the high-performance computing: Grid Computing, Parallel Pipelining, Parallel Partitioning and Massively Parallel Processing MPP.

Grid Computing

Grid Computing allows to break analytic operations into subtasks and to distribute the workload across hardware resources. In architectures of this technique, the data operations are processed in parallel on network cluster of servers, and then the results are pulled together for consumer (another process or view for end-user). Generally, resources are loosely coupled so computing power of system can be expanded by adding nodes, thus workload is increased dynamically. This approach is called scale-out and is being achieved economically by creating a logically shared address space that is physically implemented on the grid in a networked cluster of VLM, multi-core blade servers.

Grid computing with **scale-out** capabilities combined with **scale-up** capabilities enabled by in-database and in-memory processing options makes high-performance computing for advanced analytics complete. This system can effectively address the range of performance challenges that are inherent in deploying analytic models for scoring and in enabling the consumption of advanced analytics by users of BI systems and other applications.

A **message-passing interface** (MPI), which is one of the components of parallel grid computing architecture, enables to program and deploy advanced analytics applications so they that can share data and work required by analytic operations.

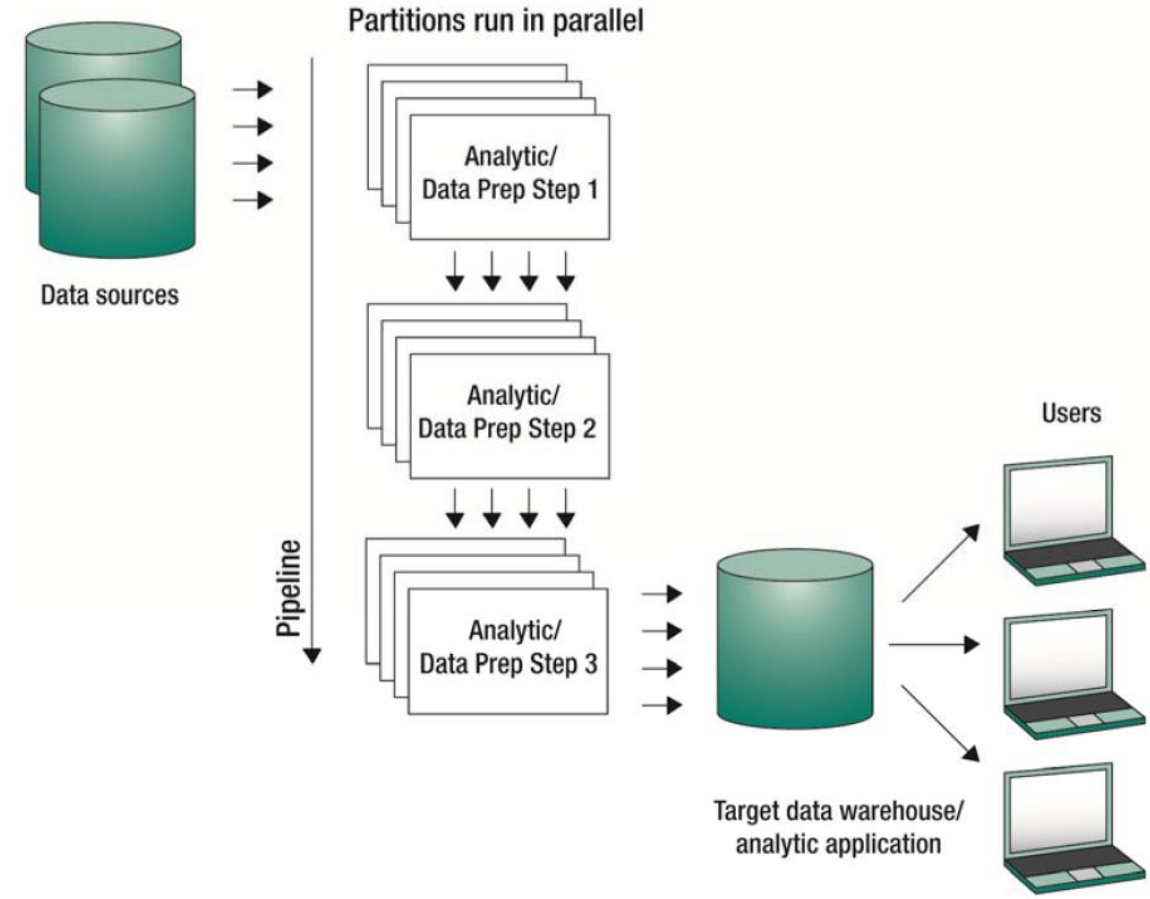


Figure 10: Pipeline and Partition Parallelism to distribute workload. [7]

Pipeline Parallelism

Pipeline Parallelism describes a multistep process, involving multiple processors, so that each step is executed on one processor (e.g. steps can be matching, loading, scoring). A sequence of steps represents the pipeline. Analytic operations would be divided into steps and pipeline parallelism allows processors to begin executing on new step instructions (of new analytical operation) even though all the steps (of previous analytical operation) in a pipeline may not yet be complete. Data, that has to wait between steps for the next one to be completed, can be placed in memory buffers.

Partition Parallelism

Partition Parallelism focuses essentially on data rather than process steps. Partitioning splits data records into subsets and push the subsets to multiple servers. The servers then work simultaneously in parallel to perform each operation. Partitions could be subsets based on data from a particular region, customer segment, and so on. Advanced analytics can take advantage of a range of partitioning methods, with the choice depending on the best way of splitting data and execute the particular type of analytics workload on each processor.

The partitioning methods split data based algorithms that include range, hash, round robin, and random partitioning.

Potentially, the combination of partition and pipeline parallelism can be used (shown at Figure 10). First, data are extracted from source systems and could be divided into partitions and run through a pipeline process for analytic operations steps, ending with the computed result data available.

4.4.2 Analytical Platforms

In contrast with data management systems focused primary on transactional processing, analytical platforms focus entirely on analytical processing at the expense of transactional processing. The breakthrough of this platform on the market started in 2002 when Netezza (now acquired by IBM) introduced an appliance based on tight integration of hardware-software database management system. This appliance was designed to support and run ad-hoc queries at outstandingly fast speeds.

Since then, many software vendors in the market made a major investment into this new analytical technology. There were another try-outs in the technology innovations from Teradata (first massively parallel database management system) and Sybase (first columnar database).

Definition

An analytical platform is a data management system optimised for query processing and analytics that provides superior price-performance and availability compared with general purpose of management systems. [6]

Technologies

Technologies behind the Analytical Platforms may diverse. The overview of these databases, systems, services or file systems is presented in Table 6. The common characteristics of all are the improvements in a price:performance ratio, availability, load times, and manageability in comparison with general rational management systems.

Technology	Details	Vendors/Products
Massively Parallel Processing (Analytic Databases)	• Row based databases • Parallel queries	Teradata Active Data Warehouse, Greenplum (EMC), Microsoft Parallel Data Warehouse, Aster Data (Teradata), Kognitio, Dataupia
Columnar Databases	• Data in columns stored • High data compression	ParAccel, Infobright, Sand Technology, Sybase IQ (SAP), Vertica (Hewlett-Packard), 1010data, Exasol, Calpont
Analytical Appliances	• Pre-configured hardware-software systems • Query processing and analytics with tuning	Netezza (IBM), Teradata Appliances, Oracle Exadata, Greenplum Data Computing Appliance (EMC)
Analytical Bundles	• Pre-defined hardware-software systems • Certified to meet performance criteria • Configuration needed	IBM SmartAnalytics, Microsoft FastTrack
In-Memory Analytical Databases	• Load data into memory to execute queries and analysis	SAP HANA, Cognos TM1 (IBM), QlikView, Membase
Distributed File-Based System	• Storing, indexing, manipulating, querying • Unstructured, semi-structured	Hadoop (Apache, Cloudera, MapR, IBM, HortonWorks), Apache Hive, Apache Pig
Analytical Services	• Private hosted • Public Cloud based	1010data, Kognitio
Non-relational Databases	• Optimised for querying unstructured and structured data	MarkLogic Server, MongoDB, Splunk, Attivio, Endeca, Apache Cassandra, Apache Hbase
CEP/Streaming Engines	• Ingest, filter, calculate and correlate volumes of discrete events • Apply rules that trigger alerts (when conditions are met)	IBM, Tibco, Streambase, Sybase (Aleri), Opalma, Vitria, Informatica

Table 6: Diversity of technologies for Analytical Platforms. [6]

Techniques

As described above, the concepts and architectures of the analytical platforms vary remarkably. Nevertheless, there are techniques in common that improve price: performance ratio, for instance Massively Parallel Processing (highly scalable), balanced configuration, storage-level processing, columnar storage and compression, memory, query optimiser or plug-in analytics. These techniques improve in particular speed of query processing and expanding footprints of used memory to provide cost effective solutions.

Massively Parallel Processing relies on multiple nodes, each contains own CPU, memory and storage and connected to a high-speed backplane. MPP systems are highly scalable, since one simply add nodes to increase processing power.

Analytical platforms **optimise the configuration** of CPU, memory and disk for query processing. Analytical appliances has wired configuration into the system, conversely analytical bundles or databases allows administrative users to configure underlying hardware in order to match unique requirements per application.

Innovative approach with breakthrough device was to move some database functions (data filtering) into hardware of a storage system using FPGA (field-programmable gate array). This technique, **storage-level processing**, reduces the amount of data that the DBMS has to process, which significantly increases query performance.

Another technique dealt with storing data in columns, not rows. Since most queries ask for a subset of columns in a row (for instance when applying the "where" clause) rather than all rows, these **columnar storages** minimize the amount of data that needs to be retrieved from disk and processed by the database, thus accelerating query performance. Additionally, since data elements in many columns are repeated (containing the same value), storing data in columns can eliminate duplicates and **compress data** volumes significantly. This technique enables more data to be loaded into memory that speeds processing and minimizes the volume of disk required to store data.

Some platforms use memory caches either for storing all data **in-memory** or **cached** the recently queried results. With the growing affordability of memory and the widespread deployment of 64-bit operating systems, which relax constraints on the amount of data that can be held in memory (due to addressing).

Query optimizer is another technique, which is worth of investing the time and resources, researching the ways to enhance the query execution into various workloads. Many analytical platforms offer built-in support for complex analytics, which include complex SQL (e.g. correlated sub-queries) or library of analytical routines (fuzzy matching algorithm, market basket calculation, support for MapReduce, etc.) as **plug-in analytics** to the database.

Deployment

In terms of products, the analytical platforms can be deployed as databases, appliances, offered as services or file-based systems.

Analytical Database can be described as a software only solution of analytical platform that runs on various hardware purchased separately. Before using the system, it has to be installed, configured and tuned, including the database engine. MPP database, columnar database, in-memory database qualify into this deployment option.

Analytical Appliance is hardware-software combining solution designed to support ad-hoc queries and another analytical processing. This tight integration of hardware and software using often proprietary components optimising performance and minimising need of adjustments and tuning. Analytical bundles, which are subset of this product option, among the other appliances give customers more flexibility to tune and configure the system, but scarify deployment itself.

Analytical Service enables customer to build the system in an off-site hosted environment or public cloud. This eliminates up-front capital expenses and diminishes maintenance.

File-based Analytical System refers in general to Hadoop, but can refer to NoSql or non-relational databases as well. This deployment option is used to store and analyse large volumes of unstructured data and does not require an up-front schema design.

4.5 Discussion

Analytical platforms of HPA vary among the vendors and usually are implemented as proprietary solutions. Specialisation of platforms is based on business requirements and business applications. Also various technical drivers affect the implementation of platform, which can be for instance number of clients/subscribers in terms of applications and users, data volumes, type of data, fit in the global architecture of BI solution (e.g. staging area, data-warehouse, dependent/independent data marts, ad-hoc research and prototyping facility).

Speaking about trends, information silos (enterprise data-warehouse) are moving towards pooled resources where data are separate according to the data importance (priority processing), data format (structured, semi-structured, unstructured).

Infrastructure and architecture of data analysis and management system are moving from being performance tuned towards being linearly scalable (in linkage with distributed parallel processing, grid computing, and in-memory analytics). Scalability motivates another paradigm that shifts on premises deployment towards hybrid (traditional and appliances oriented deployment), where scalability can be adjustable with private cloud.

5 Assignment Specification

The practical part of the thesis starts with the specification that determines a goal of the assignment, a definition with including tasks, a scope and limitations of the assignment.

The outcome of the experimental assignment includes analytical processing of huge dataset exploiting an analytical platform from SAS Institute. The experiment demonstrates analytical processing for selected HPA methods that are discussed in the theoretical part.

5.1 Goal

The goal of the practical assignment is to implement the whole analytical cycle on analytical platform of HPA in order to perform the tests on different datasets. The analytical cycle is comprised with multiple steps:

1. Acquisition and Preparation of Data
2. Loading into In-Memory Storage
3. Profiling of Data
4. Definition of Metadata
5. Exploration Analysis
6. Creating Complex Reports
7. Presentation of Outputs via web and/or mobile device

First, the **analytical cycle** starts with acquisition and preparation of data that afterwards are loaded into in-memory storage (in case of HPA In-Memory Analytics). Data profiling gives an overview and statistical information about data in rows and/or columns (typically helpful for initial familiarisation with extremes and values in a concrete dataset, histogram of data). In the next step, dataset is enriched with metadata (e.g. adjusting naming used in dataset, dimensions and hierarchy modelling, etc.).

The core of the analytical cycle is an exploration analysis. In this assignment, concept of the exploration analysis covers advanced visualisation techniques for exploration methods, such as classical reporting (dashboards, charts, complex reports etc.), multidimensional analysis (drill-down/up, slice and dice, pivoting, etc.), statistical and special methods (forecasting, correlations, regressions, etc.), geographical analysis and/or combinations of them (e.g. reporting + forecast). Visualisation of results can be considered as means of the presentation of outputs.

A test procedure includes observation and measurement of the response time of system, which is necessary for calculation and visualisation performed on HPA platform, and comparison of the response times for various exploration methods (analytical scenarios) on different datasets.

Last, a discussion of analytical steps, outputs, observations and measurements in context of HPA In-Memory is held as a conclusion of experiments with analytical platform that pinpoints advantages and differences against the traditional OLAP analysis.

5.2 Definition of Assignment

The definition of experimental assignment consists of the implementation of an analytical cycle for two tasks and summarisation and evaluation of achieved results as the output of those tasks. Complete set of visualisations and reports should lead into the final analytical product in order to test a response time for given of analytical platform. By completing both

tasks there is an opportunity for discussion about the assets and contribution of HPA In-Memory architecture to appropriately support (or not) different analytical methods or operations.

Task A

Implement an analytical cycle (mentioned above) on the infrastructure of local system at SAS Institute CR, s.r.o.. Particularly, implementation should include preparation of dataset from automotive domain oriented for car sales and marketing purposes. Dataset should be scalable for sampling from 100k to 1M records (same structure, different amount, and different distribution). Further, it includes loading the dataset into environment of in-memory storage of analytical platform. Implementation should proceed with set of analysis utilising multiple analytical methods or operations and test the response time on various datasets.

Task B

Implement an analytical cycle on the infrastructure of central system at SAS Institute Corp, Inc.. Particularly, implementation should include selection of existing already prepared dataset (on central shared system) that should be scalable for sampling from 100M to 1G. Likewise task A, task B includes applying the set of analytical methods, operations and reporting procedures.

5.3 Scope

The scope of the assignment is to complete the defined tasks (see above). Especially for the scope of task A, the whole analytical cycle should be implemented for dataset scaled for sampling from 100k till M records. Alternatively for the scope of task B, once the dataset is available, exploration analysis and output (reporting, testing) should be implemented. In the second case, the selected dataset should be scalable for sampling from 100M till 1G of containing records. The analytical cycle for both tasks should be implemented on software and analytical platform form SAS that utilises HPA approach with In-Memory Analytics.

5.4 Limitations

There are some limitations regarding to the practical assignment. First, implementation of goals and tasks of the experiments are performed on one analytical platform from vendor SAS Institute. These technological platform represents a proprietary solution implementing the idea of high-performance analytics. Due to the deployment requirements (hardware, configuration), licensing, etc., the experiments are held on the in-house infrastructure and environment of SAS Institute. In addition, the limited access has been provided by supervisor of this thesis who is employed at the institute.

Second, provided infrastructure and environments contains installed and configured product SAS Visual Analytics. The analytical platform of this product is based on solution that involved realisation of the concepts In-Memory Analytics, Grid Computing and combining with File-Based Distributed System (Hadoop) for balancing, replicating, distributing the data. Downside of given solution from perspective of the experiments is that In-Database Analytics are not part of provided system. Therefore the full demonstration and comparison is of all HPA concepts are not achievable with this analytical platform.

6 Environment and Methodology

This section illustrates the environment where the experiments and tasks are implemented. Further, it describes process and procedures to be followed in order to achieve the implementation of analytical cycle. Next, it describes methodology to be applied in comparing of obtained results.

6.1 Infrastructure

There are two environments disposable at SAS Institute. A local system in the infrastructure located at SAS Institute CR s.r.o., and a global system in the infrastructure located at SAS Institute Inc. that is shared among offices. Both system configurations of disposable systems are summarised in Table 7.

	Local System Specification	**Global System Specifications**
Server	HP Proliant DL380p Gen8	Cisco/B200M3
Processors	2 x Intel® Xeon® E5-2640 (6 core, 2.50 GHz, cache 15MB, 95W)	16 x Intel® Xeon® E5-2680 (8 core, 2.7 GHz, cache 20MB, 130W)
Cores total	24 (leverage by using technology with 2 hypervisors)	256 (leverage by using technology with 2 hypervisors)
RAM	256 GB	16 TB
Available storage area	2.4 TB	27 TB

Table 7: Configuration of systems with a deployed analytical platform

6.2 Environment

The environment and technical platform of SAS Visual Analytics is illustrated with details of architecture overview and technologies.

6.2.1 (HPA) Method Classification

According to the classification introduced in Section 4.3, SAS Visual Analytics could be classified as follows. By the tasks, as analytical platform it is not narrowly specialised, it covers the reporting, exploration and fits the advanced analytical usage (forecasting, correlations, and regressions). By the dimensions (of Big Data), it supports the most the data volume (data loaded in operational memory - In-Memory-Analytics), velocity and value. By the resources, it can be primary classified as in-memory (utilising the operational memory) and parallel processing (utilising the grid computing on blade servers).

6.2.2 Architecture and Technologies

This section contains details about the analytical platform – under the hood of SAS Visual Analytics. The architecture high-level overview is illustrated in Figure 11. Illustrations shows that the solutions is configured in eight blade commodity cased chassis where one represents a management and servicing server (containing middle tier, compute tier, metadata server, workspace server) and cluster scaled-out into seven nodes.

Several technologies and solutions are used for one node, such as SAS LARS Analytical Server (facilitating the in-memory processing), SAS Visual Analytics Hadoop (containing the data balancing, replication and block distribution across nodes), TKGrid (allows for the execution of partitioned analytics to run on a cluster), and Message Passing Interface (allowing the nodes of the cluster to communicate with one another).

The architecture of the SAS Visual Analytics is described in more details in Appendix A.

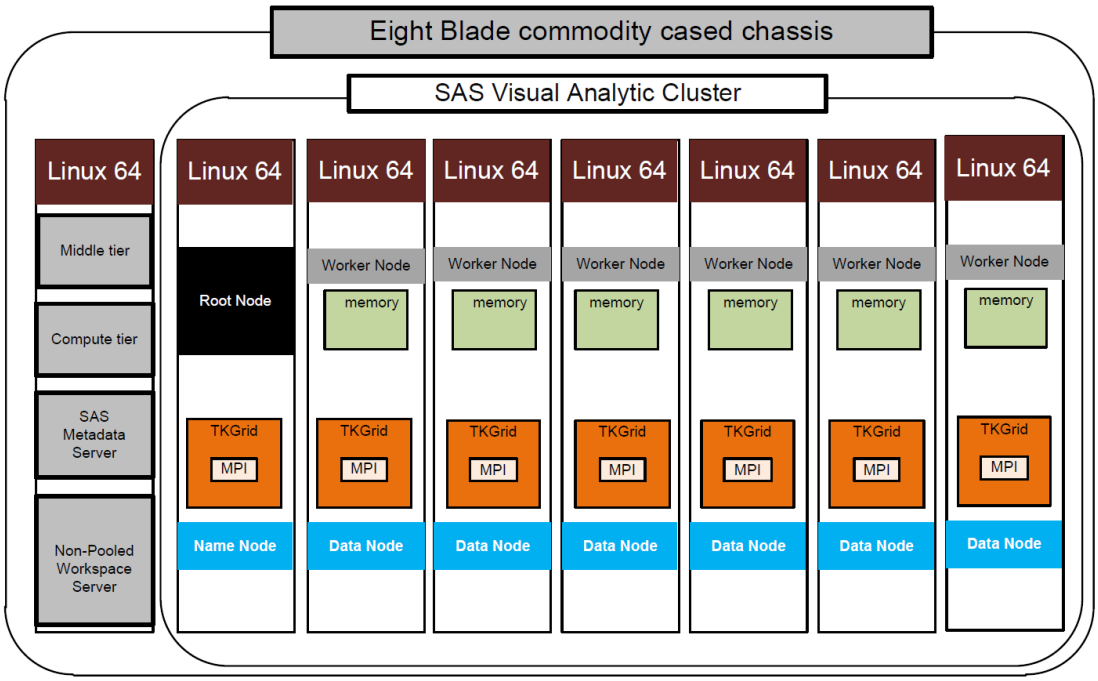


Figure 11: Architecture of Visual Analytics Cluster

Visual Analytic Clients

Visual Analytics works with web based client access via IE and or Firefox. The client access contains following client tools. Visual Analytics Hub (VA-H) is the central entry point for access to all role based view and functions. Visual Analytics Explorer (VA-E) is used for ad-hoc data discovery and visualization to allow users to explore and analyse their data. Visual Analytics Designer (VA-D) is for the creation of reports and dashboards for a mobile client. Mobile BI is the native iPad application available from the Apple iTunes store that allows viewing of reports created through VA-D. Data Preparation is the set of capabilities that IT will use to manage users, blacklist mobile devices, monitor servers etc. There are also data-related administration tasks to load data into the SAS LASR Analytic Server instances, perform joins and create calculated columns.

6.3 Process and Evaluation Technique

This section is oriented on process of fulfilling the goal and implementing defined tasks. Afterwards, the atomic **Analytical Operations** are defined from which the **Analytical Scenarios** may be composed. Last subsection described the criteria of performance testing of applied Analytical Scenarios.

6.3.1 Process

The process of fulfilling the goal and implementing task is defined in following steps:

1. Acquisition and preparation of data
2. Load into in-memory storage
3. Profiling of Data
4. Definition of Metadata
5. Exploration Analysis
    a. Define the Analytical Scenarios
    b. Constructing the Analytical Scenarios (by using SAS Visual Analytics)
    c. Test the performance for Analytical Scenarios

6. Complex Reports Creation
7. Presentation of Outputs (via web and/or mobile device)

Profiling of Data should unveil the number of records and per minimum and maximum value, number of distinct values, and histogram of data distribution. The step Definition of Metadata includes adjusting of the loaded data model by modifying displayed names (aliases) and data formats, by creating hierarchies out of Categories (equivalent with dimensions in OLAP), by constructing calculated variables (e.g. Satisfaction of Customers).

Exploration Analysis requires first to compose a scenario from analytical operations (e.g. drill-down, forecasting see 6.3.2) and completed them with visualisation techniques. Next, the scenarios should be constructed in environment of an analytical platform SAS Visual Analytics. Further, the performance of application of these analytical scenarios should be tested (see 6.3.3).

6.3.2 Operations

This section determines and describe the operations that are related to data analysis. Operations are split into classical explorative analysis (mostly applied in reporting), multidimensional analysis (equivalent to the OLAP explorations), and advanced statistical analysis. In following text, a short description is given for each of these categories and eventually for combinations of them.

Classical Analysis/Reporting

Classical analysis and reporting contains simple visualisations, e.g. tables showing raw data, line/bar charts showing the relationship between variables, scatter plots showing dependencies between variables. Calculating percentage or summarising values (pivoting, sort, count, total, average, etc.) can be part of classical data analysis. Geographical Analysis, which displays the data as an overlay on the map, can be regarded as a special case of classical analysis.

Analytical operations for these types of analysis are typically not interactive, thus they are trivial and limited in the displaying the data.

Multidimensional Analysis

This type of analysis includes aggregations for categories or levels of hierarchies (in case of Visual Analytics, the aggregations are not stored but calculated on-the-fly); multidimensional explorations (including at least 3 measures) with multidimensional aggregations; calculated variables (measures or categories); multidimensional tables (in case of Visual Analytics, they are called Crosstabs) that enables to display data with more measures and categories and thus used the concept of aggregations directly for each cell.

Analytical operations for these types of analysis are drill-down, drill-up, drill-across (all previous: navigations in hierarchies), slice and dice (simple filtering by existing values), filtering on custom conditions, operators and expressions.

Statistical Analysis

In the scope of this experiment, there are three types of statistical analysis recognised.

Correlation identifies the degree of statistical relationship between measures. Strength of a correlation is described as a number between -1 and 1.

Trend Line plots a model of the relationship between measures. There are many types of trend line, including linear trend (using a linear regression algorithm), quadratic trend (trend with a single curve), cubic trend (trend with two curves), and penalized B-spline (smoothing spline that fits the data closely).

Forecasting estimates future values for data based on statistical trends. Forecasting is applicable only for subset of data that are combined with date or time. The forecast models (Damped-trend exponential smoothing, linear exponential smoothing, Winters method, etc.) are used for estimations. The x% confidence interval consist of a range of values that holds the true value of desired forecast model with probability equal to x%.

Additionally another statistical operations can be calculating percentile (25^{th} and 75^{th}), median, mean (usually any other operation that is not supported directly by SQL querying), etc.

Combinations

Combinations of all above lead to creating business reports to reflect the real life usage, therefore it is determined as a separate category. Reports contains multiple visualisations (charts, plots, gauges, tables, multidimensional tables, matrices, maps), moreover they are required to be connected between each other when user applies simple selection, or any analytical operation that involves re-calculation and rendering for particularly selected item (likewise master-detail view concept). In the scope of this experimental assignment, the following combinations can be considered: reporting + forecast analysis, reporting + multidimensional analysis, statistics + multidimensional analysis, or any reasonable combination.

6.3.3 Evaluation

A Testing Procedure is focused on a response time of Analysis Scenarios running on analytical platform. Scenario may contain one, or combination of operations (see Section 6.3.2). Evaluation includes steps:

1. Measure a response time of Analytical Operations that takes to calculated the output for selected Analysis Scenario
2. Compare a response time for different scenarios
3. Compare a response time for different samples of datasets

Response time can be defined as a time that is need for calculations of underlying data and partly including the time needed for visualisation of data. For instance, in case when geographical map is used the response time includes pre-calculation of geographical data in combination with underlying business data.

For comparison of different scenarios or analytical operation there is an idea to split them into related groups in terms of required calculation, e.g. basic, multidimensional, advanced statistical, and combinative.

Discussion should be held in order to assess the convenience of the tested analytical platform for different types of analysis.

7 Implementation

This section covers the implementation of the experimental assignment. In particular, it covers the used dataset and tools, describes the outputs and process that is involved in order to obtain them.

7.1 Datasets in Details

There are two dataset for the experimental assignment. A dataset of the Car Sales and Marketing has been created and populated, whereas a dataset of the Toy Manufacturing Corporation has been adopted from testing environment in SAS Institute Inc.

Dataset of the Car Sales and Marketing (Cars)

A dataset used for task A is a model of automotive sales and marketing data. Model contains four main entities: Car Dealer, Customer, Vehicle, Campaign, and Event. Further, the Car Dealer entity contains information about location, the Customer entity contains information about personal details (name, birthday, birthday code, gender, etc.), living location (region, city, postal code, etc.), the Vehicle entity contains information about type, class, and a production date. The Campaign entity contains start date, end date, and success of the campaign, number of events within campaign. The Event entity contains information about the type of event (purchase, service), financial value, and date.

The data model is de-normalised and overview of the entity attributes is given in Appendix B in Table 14. With de-normalisation, structure of entities with attributes is transformed into variable (each attribute becomes variable). Each variable of entity conveys information about its data type, type of variable, and unique count (output of a data profiling).

There are two types of variables, Category (Element of Dimension in OLAP terminology) and Measure (Measure in OLAP terminology). Further, there are two common variables, which are typically predefined and supported by exploration techniques such as time variable, or geographical variable, due to common use among different data models and special way of visualisation. Another special variable is calculated variable, the value is calculated on-the-fly rather than being stored.

The Categories can be organised as Hierarchy that defines set of parent-child relationship designing that parent summarises its children. Parent elements can further be aggregated as the children of another parent (however difference with OLAP in this case is that HPA does not store aggregated values for performance optimisation).

Reasonable hierarchies, in the model of automotive sales and marketing, are summarised in Table 8.

Name	Relationship
Vehicles	Class of vehicle – Type of vehicle – Id of Vehicle
Event Monthly	Month of Event – (Full) Date of Event
Event Yearly	Year of Event – Month of Event – (Full) Date of Event
Location of Dealer	Dealer's Region – Dealer's Canton – Dealer's City
Location of Customer	Customer's Region – Customer's Canton – Customer's City

Table 8: Hierarchies for a Car Sales and Marketing Dataset

Dataset of the Manufacturing Corporation (ToyCorps)

A dataset used for task B is a model of the corporate manufacturing and production data. Model contains three main entities, such as Product, Unit, and Facility. Further, the Product

entity contains information about a material cost, a final price, and a product quality. The Unit entity contains information about its production capacity, lifespan, and reliability. The Facility entity represents a branch of corporation and a collection of Units. In particular it contains information about locations and expenses.

Again, the data model is de-normalised and overview of the entity attributes is given in Appendix B in Table 15. There are two common variables, which are typically predefined and supported by exploration techniques such as time variable, or geographical variable, due to common use among different data models and special way of visualisation. Another special variable is calculated variable, the value is calculated on-the-fly rather than being stored.

Reasonable hierarchies, in the model of automotive sales and marketing, are summarised in Table 9.

Name	Relationship
Date	Date by Year - Date by Month - Date
Location of Facility	Facility Region - facility State - Facility City
Facilities	Facility Type - Facility - Unit
Products	Product Brand - Product Line - Product - Product Description - Unit

Table 9: Hierarchies for a Toy Manufacturing Corporation Dataset

7.2 Tools in Details

The goal and tasks of the experimental assignment are achieved and implemented by using Visual Analytics Tools (VAT). In this section, the description of those tools is given together with visualisations that utilises the procedures defined in Section 6.3.2. VAT works with objects such as visualisation, exploration, and report.

Visual Analytics Explorer (VA-E) enables to explore data sources by using **visualisations** such as charts, histograms, and tables, alternatively with forecasting and correlation methods. An **exploration** contains instances of visualisations, metadata, data settings and filters). Alternatively, exploration and results of exploration can be deployed as **report**. Report can be further refined and customised in Visual Analytics Designer (VA-D). Last tool is Visual Analytics Viewer (VA-V) that displays reports.

A data model can be created from categories and measures (see Section 7.1). However there are special features available that can tackle aggregated measures, percentage measures, calculated measures, geography data items (containing latitude and longitude).

7.2.1 Visualisation Types

Available visualisation types that enables to explore and examine data in Visual Analytics are summarised in Table 10.

Visualisation	Utilisation
Table	raw data, columns arrangement, sorting
Crosstab	data for intersection of hierarchy nodes/levels, or categories, cells shows aggregated data (similar to reporting on OLAP – multidimensional table)
Bar Chart	comparing data that is aggregated by the distinct value of a category (horizontal, vertical, grouping in lattices)
Line Chart	data trends over time (grouping in lattices)
Scatter Plot	useful for relationship between numeric data items, correlation/regression analysis

Bubble plot	useful for relationship between at least three measures (2 axes, size of marker of plot), animate bubble plot over time
Histogram	distribution of value for single measure, percentage count
Box Plot	distribution of values for a single measure using box and whiskers (size of plot: range of values that are between 25 a 75 percentile, dost: mean, horizontal line: median)
Heat map	distribution of values for two measures using table with collared cells, without the another measure for colour, than represents the frequency of each intersection of values
Geo Map	bubble markers for data interpretation as overlay on a geographic map
Treemap	data as a set of rectangles (tiles) representing category value, size of tiles represents frequency, or value of measure, colour indicating additional measure
Correlation Matrix	displays the degree of correlation between measures as a series of coloured rectangles, each colour indicates the strength of correlation (using Pearson's product-moment correlation coefficient: weak <0.3, moderate 0.3-0.6, strong 0.6<)
Fit Line	a model of the relationship between the variables (linear fit, quadratic fit, cubic fit, and penalized B-spline)

Table 10: Visualisation Types available in Visual Analytics

7.2.2 Data Analysis

From advanced data analysis Visual Analytics offers forecasting (input data and duration of prediction are customisable), calculation of 25th and 75th percentile, median, mean (these calculations are utilised in Box Plot visualisation).

Correlation feature is available together with trend line (utilised in Fit Line visualisation) that uses a linear regression algorithm, quadratic trend (trend with a single curve), cubic trend (trend with two curves), and penalized B-spline (smoothing spline that fits the data closely). Correlation Matrices displays the degree of correlation between multiple intersections of measures as a matrix of rectangular cells.

7.2.3 Outputs

At Visual Analytics there are multiple options how to deliver output. On-line Analysis in an option where user can interact with visualisation and explore data. Another option is reporting that involves deployment of explorations or exploration results as reports for further access. Similar option is deployment of mobile reports that can be accessed either by smartphone or tablet (connectivity required). Last option is represented by exportation of visualisation or reports into static or off-line mode.

7.3 Process Realisation in Details (Methods)

Following defined process in Section 6.3.1 and defined analytical operations in Section 6.3.2, this section is focused on realisation of the process. Therefore the analytical scenarios are determined as the interesting combinations of analytical operations. List of scenarios is summarised in Table 11.

Scenarios are later applied on each pre-defined sample of datasets in order to get the outputs and the observation results. Analytical Scenarios are referenced using code in evaluation of results and discussion later in Section 8.

Code	Description	Analytical Operations
Import	Import records into memory	
Line	Line Chart	
LineDrillDown	Line Chart - drill down	Drill down
Forecast3m(6)	Forecast – 3 measures – 6 months	Forecasting
Forecast3m(18)	Forecast – 3 measures – 18 months	Forecasting
Forecast1m(6)	Forecast – 1 measure – 6 months	Forecasting
Forecast1m(18)	Forecast – 1 measure – 18 months	Forecasting
Bar	Bar Chart	
BarDrillDown	Bar Chart – drill down	Drill down
BarChart3d	Bar Chart – 3 categories	
BarChart3dDrillDown	Bar Chart – 3 categories – drill down	Drill down
GeoMap	Geography Map	
GeoMapDrillDown	Geography Map – drill down	Drill down
BoxPlot	Box Plot	Mean, Median, Percentile
BoxPlotLattice	Box Plot – lattice	Mean, Median, Percentile
BoxPlotLattice2m	box Plot – lattice – 2 measures	Percentile, Median, Mean, Multi-dimensions
BoxPlotLatiice2mDrillDown	box Plot – lattice – 2 measures – drill down	Percentile, Median, Mean, Multi-dimensions, Drill down
Corr24md	Correlation – 24 measures/categories	Correlation (various methods), Multi-dimensions
CorrLinear	Correlation – linear	Correlation
CorrQuadratic	Correlation – quadratic	Correlation
CorrPSPline	Correlation – PSP line	Correlation
HeatMap	Heat Map	Multi-dimensions
BubblePlotAnim	Bubble Plot – animation	
Crosstab3d5m	Crosstab – 3 categories – 5 measures	Multi-dimensions
Crosstab3d5mDrillDown	Crosstab – 3 categories – 5 measures – drill down	Multi-dimensions, Drill down

Table 11: Description of the Analytical Scenarios

7.4 Documentation

This section documents the sample of data for used datasets, additional metadata for datasets, and the outputs including visualisations and reports.

Sample of Dataset	Number of Rows	Number of Columns
Car Sales and Marketing (sample 1)	500 000	52
Car Sales and Marketing (sample 2)	1 015 046	52
Toy Manufacturing Corporation (sample 3)	1 158 601 468	45

Table 12: Number of records and columns in the dataset samples

7.4.1 Sample of Datasets

Statistics about the size of datasets are summarised in Table 12. Datasets are introduced in Section 7.1, described in Appendix B. Sample of dataset can be found in Attachments.

Details about the data in ToyCorps dataset are illustrated in Figure 12, including min/max values for measures, and optionally histogram of values per specific measure.

Figure 12: Details of measures in ToyCorps dataset.

7.4.2 Sample of Metadata

In the Figure 13, the example of hierarchy is shown. Hierarchy for Location of Dealer is composed from categories Dealer's Region and Dealer's Canton and Dealer's City.

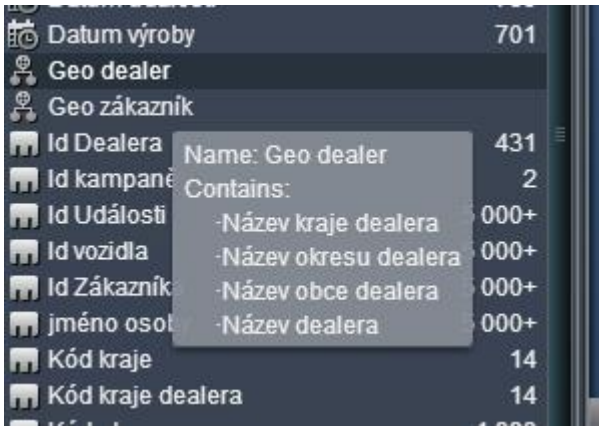


Figure 13: Details of hierarchy for a Location of Dealer

7.4.3 Process

In the Figure 14, the step of importing data is shown. Particularly, the import of data for Dataset A (Car Sales and Marketing Data).

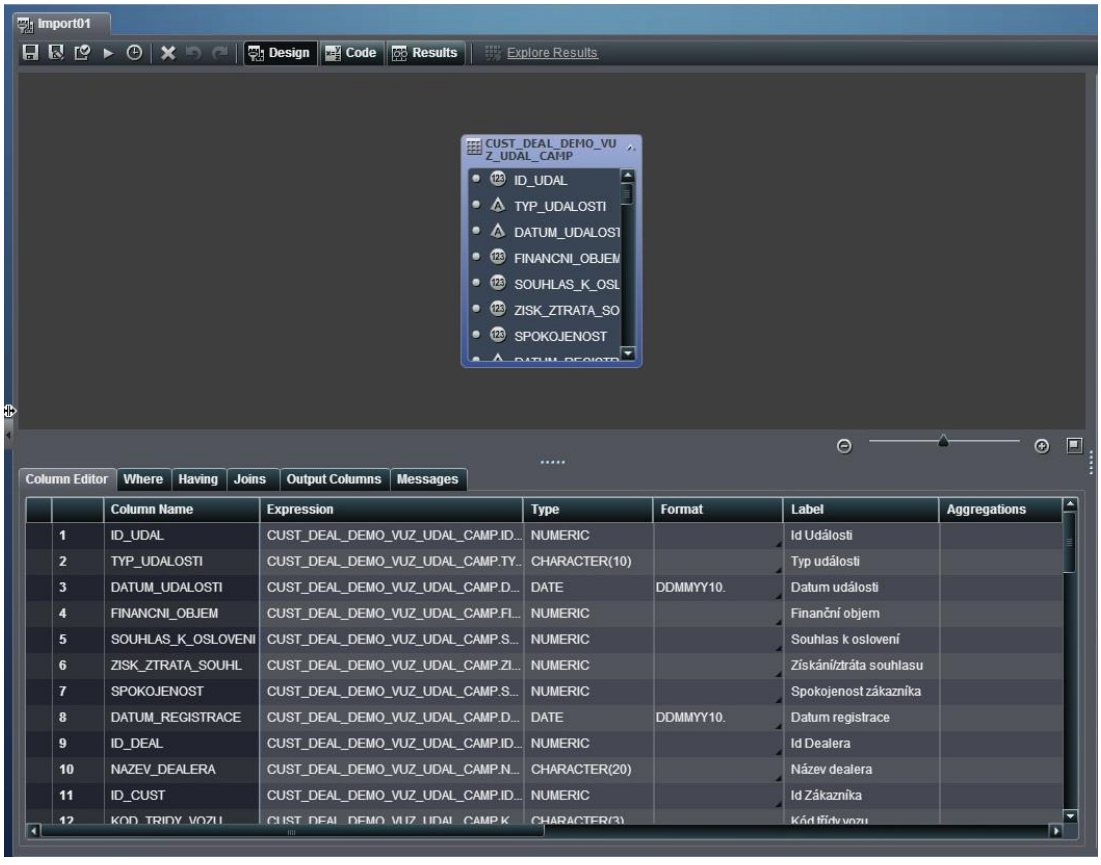


Figure 14: Import of Dataset A (Car Sales and Marketing Data)

Visual Analytics Explorer (VA-E) is one of the used tool during the process of implementation. The environment of VA-E is shown in Figure 15.

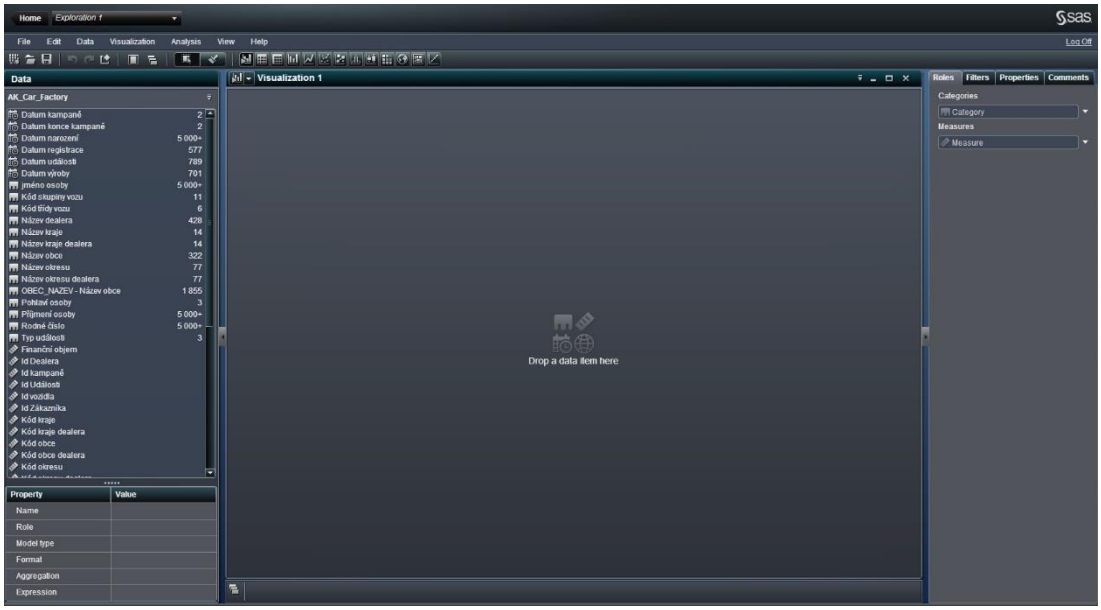


Figure 15: Visual Analytics Explorer for creating analytical scenarios using visualisations.

7.4.4 Sample of Output

There are a few visualisations presented in this section. The complete set of visualisations and reports can be found in the attachment (enclosed CD).

Explorative visualisations

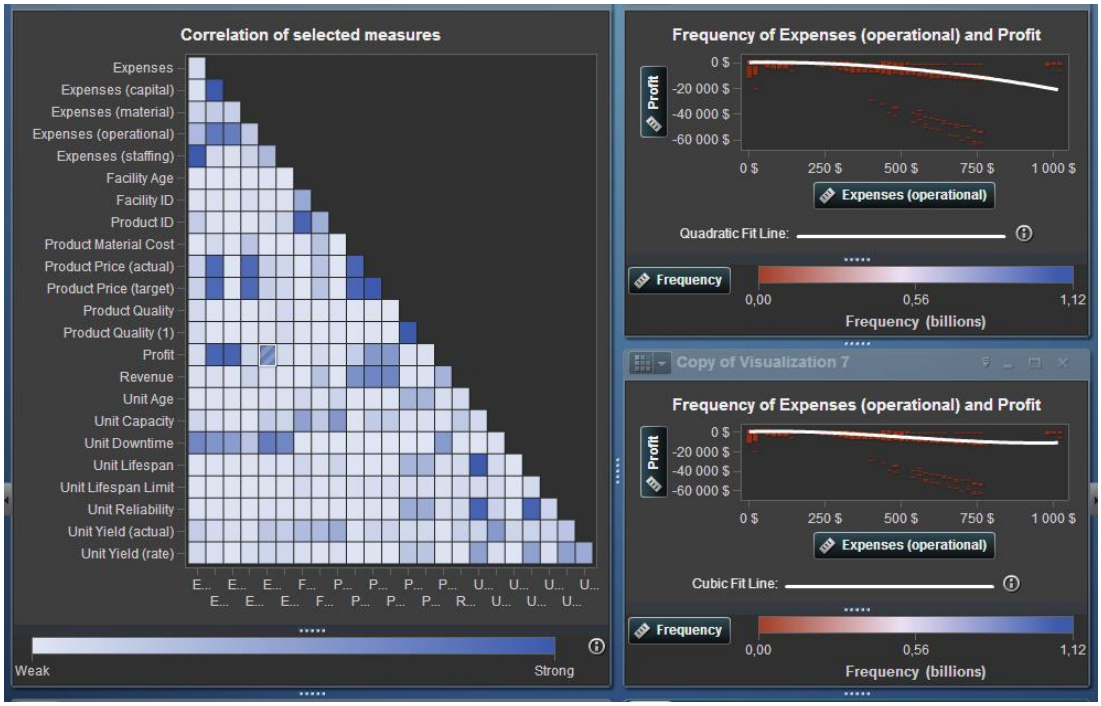


Figure 16: Correlation of measures with detailed trend-lines for specific pairs.

Figure 17: Forecasting

Reporting

There are a few reports presented in this section. The complete set of reports can be found in the attachment on enclosed CD. The reports are compositions of interconnected visualisations, where user action, for instance analytical operation drill down, triggers recalculations for all visualisations so the data are display for the same context.

Figure 18: Report with Treemap and Charts visuallisations.

Figure 19: Report with multiple visualisations.

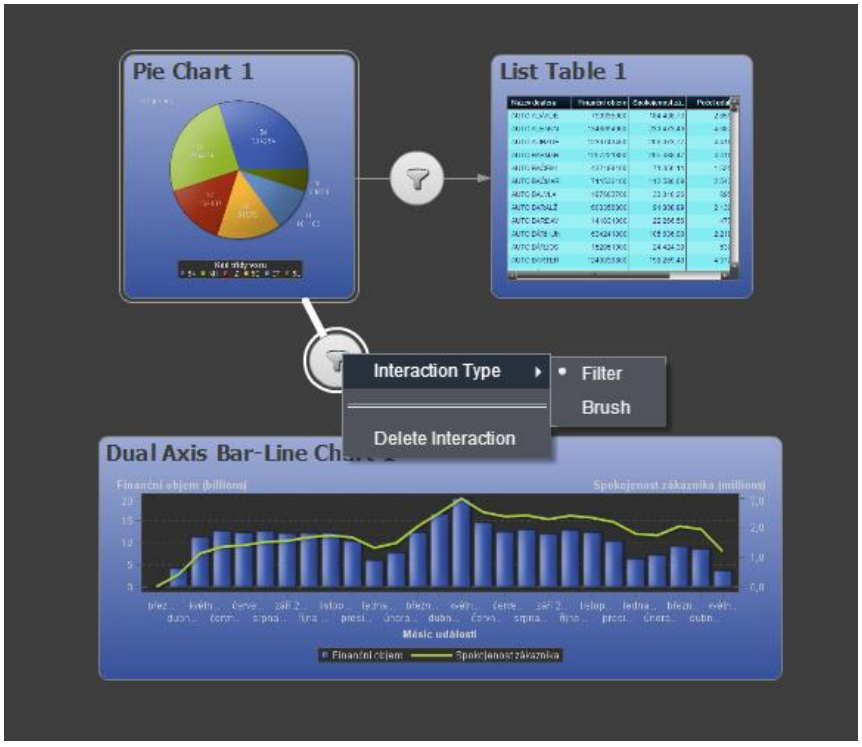


Figure 20: Interconnectivity in reports with multiple visualisations.

8 Results

This section unveils the results obtained from implementations of analytical cycle, as well as from analytical scenarios that generated outputs. Outputs are based on calculations on in-memory platform and visualisations via Visual Analytics Explorer (VA-E).

8.1 Evaluation

Reminding, response time can be defined as a time that is need for calculations of underlying data and partly including the time needed for visualisation of data. For instance, in case when geographical map is used the response time includes pre-calculation of geographical data in combination with underlying business data. The measures of response time for particular analytical scenarios are displayed in Table 13.

Now, let's create groups by type of calculations involved when processing outputs. There are **basic calculations**, **explorative operations** (OLAP equivalent), **advanced statistics** and **combinations** of previous (see the classification in Section 6.3.2). Once the calculations and/or analytical operations are sorted out the differences (if any) among them are unveiled and discussed.

Code	Dataset Sample 1	Dataset Sample 2	Dataset Sample 3	Note
Import	2min	2min	n-a	
Line	1s	<1s	1s	
LineDrillDown	<1s	<1s	<1s	
Forecast3m(6)	2s	3s	3s	
Forecast3m(18)	2s	3s	3s	
Forecast1m(6)	<1s	0.5s	1s	
Forecast1m(18)	<1s	1.5s	1.5s	
Bar	<1s	<1s	<1s	
BarDrillDown	<1s	<1s	<1s	
BarChart3d	5s	5s	4s	
BarChart3dDrillDown	5s	5s	4s	
GeoMap	<1s	<1s	<1s	
GeoMapDrillDown	1s	<1s	1s	
BoxPlot	<1s	<1s	<1s	
BoxPlotLattice	1s	<1s	1s	
BoxPlotLattice2m	2s	1s	1s	
BoxPlotLatiice2mDrillDown	1s	<1s	1s	
Corr24md	0.5s	2s	13s	276x linear regression
CorrLinear	1s	1s	3s	
CorrQuadratic	1s	1s	1s	
CorrPSPline	1s	1s	1s	
HeatMap	<1s	<1s	1.5s	
BubblePlotAnim	<1s	1s	3s	
Crosstab3d5m	2s	2s	7s	
Crosstab3d5mDrillDown	1.5s	1.5s	15s	

Table 13: Response times for given Analytical Scenarios on different datasets

The values of response time in Table 13 imply that basic calculation (Line, LineDrillDown, Bar, BarDrillDown, GeoMap, GeoMapDrillDown) have been processed within one second for all datasets. Let's take these values as base for comparison with other groups.

Explorative analysis, or better multidimensional analysis (Crosstab3d5m, Crosstab3d5m-DrillDown, HeatMap, BubblePlot), that typically involves the on-the-fly calculation of aggregated values is fast enough to compete with basic calculation. A slightly bigger response time is noticeable with the largest dataset containing billion of records. Although the crosstab with 5 measures and 3 categories takes much longer for the largest dataset due to the complexity, but for user the analytical operation is still reasonably fast.

Important fact to notice is that advanced statistics and advanced analytics (Forecast3m(6), Forecast3m(18), Forecast1m(6), Forecast1m(18), BoxPlot, BoxPlotLattice, CorrLinear, CorrQuadratic, CorrPSPline, BoxPlotLattice2m-DrillDown,) have been processed almost as fast as basic calculations, however they involved more complex calculations.

For special scenario, when correlation for each pair of 24 measures and categories, it took only 13 seconds to calculate all 276 linear regression on the largest dataset. This scenario involves the most intensive calculations among all scenarios. On the other hand, drill-down operation on the crosstabs has bigger response time due to the largest amount of data that should be transferred in order to visualise it.

8.2 Discussion

This section draws the discussion in order to assess the convenience of the tested analytical platform for different types of analysis. The following paragraphs contains the interesting thought that popups during the experiment.

During the observations and measurements, the overhead of transferring data form LASR Analytics Server, transferring of metadata (e.g. maps), and rendering time on front-end could not be separated from computational time of LASR engine working with datasets. This creates the space for further investigation and future work.

On experiment, two different infrastructures have been used, one located in Czech Republic (local environment) and the other in USA (global environment). Therefore the analytical scenarios performed on global environment might experience some overhead in network latency as well.

Although the structure and content of chosen dataset are different, from perspective of required computation the datasets are not that diverse (both contain the de-normalisation of 3-4 entities, geography information, contains 45-51 columns). Therefore within the scope of this experiment the obtained results from different datasets are comparable.

In-Memory Analytics architecture proves to be most convenient for the computation intense operations (correlation, trend lines, forecasting, percentile), because all data are loaded in memory and can be directly addressed. In case of using external memory (e.g. RDBMS), even though it could be in-memory storage but located in another appliance, for these analytical operations should take (tens of times) longer due to the repetitive data transfers or I/O operations (for each calculation request coming from analytical model). As this comparison is not achievable with chosen analytical platform, it is the limitation of this experiment, yet it opens space for further investigation and future work.

Combination of multidimensional operations (e.g. drill-down) and advanced statistics (e.g. forecasting, percentile) is also suitable for In-Memory Analytics architecture as well. This combination can be barely handled by traditional OLAP systems at all, because it requires the calculation of statistics form detailed data on complete set. Using the RDBMS as storage, the solution would struggle with loading the complete set by parts into memory and performed the statistical calculations, and still applying the aggregations. The direct experimental comparison of this scenario on different platforms thrives for future work.

9 Conclusion

Since the Big Data has been continuously identified for a decade there is a lot research done in literature, white papers, and online references. In this thesis, the Big Data Phenomenon is summarised in an overview including its causalities, definition, influence and impacts. It represents a starting point and driver for High Performance Analytics in terms of raw material that contains hidden information, patterns and value. Concluding from research, Big Data, with its dynamic dimensions, should not be considered as a problem, rather than opportunity to turn it into advantage.

High Performance Analytics is extensively researched in this thesis as an approach towards handling Big Data. Due to this area it is still emerging, being refined and formalised among vendors, research on HPA is challenging in order to bring overview, classification of HPA methods and techniques (In-Memory Analytics, In-Database Analytics, and Parallel Computing), their characteristics, and appropriate usage.

HPA is driven by business world with broad requirements to compute results as fast as possible on the largest dataset. The formation HPA becomes possible with technological evolution (Very Large Memory, 64bit address, Grid Computing) and affordability of hardware (costs, price:performance indicator). For now, HPA can be seen as a solution complementary to the Business Intelligence, but highly on premises the evolution will continue further. The research could be extended to dive into HPA solutions from another vendors comparing multiple proprietary approaches in details.

Experimental assignments as it has been designed, demonstrate the implementation of HPA approach on large datasets. Different analytical operations and their combinations have been selected to demonstrate the benefits of analytical platform based on In-Memory Analytics approach. In-Memory Analytics architecture has found as convenient for the computation intense operations (correlation, trend lines, forecasting, percentile), because all data are loaded in memory and can be directly addressed.

There are several limitations of experimental assignments. For the future work, the analytical operations can be tested on analytical platforms that implement other HPA approaches, also against traditional approach in Business Intelligence (e.g. OLAP). Rarely, vendors offer the analytical platform that would implement all discussed HPA approaches and if, it is challenging to perform them on the same system infrastructure.

Overcoming the limitations, the experiments are sufficient for comparing the performance of analytical operation among each other, to identify advantages and benefits of selected analytical platform.

References

[1] High-Performance Analytics; SAS Institute - White paper; 2012.

[2] Big Data Analytics: Future architectures, Skills and roadmaps for the CIO; Philip Carter; IDC White Paper; 2011.

[3] From Big Data to Meaningful Information - Insights from a webinar sponsored by KMWorld Magazine and SAS; Conclusion paper; 2013.

[4] Big Data Meets Big Data Analytics; SAS Institute - White paper; 2012.

[5] Data Visualisation Techniques; SAS Institute - White paper; 2012.

[6] Big Data Analytics: Profiling the Use of Analytical Platforms in User Organizations; Wayne Exkerson; 2011.

[7] Seven Keys to High-Performance Data Management for Advanced Analytics; TWDI Monograph Series; David Stodder; 2011.

[8] Big Data Analytics; TDWI best practices – Fourth Quarter 2011; Report Philip Russom; 2011.

[9] Intelligence Quarterly, Second Quarter 2012; SAS Institute; 2012.

[10] Postmoderní architektura Business Intelligence; Business Intelligence Fórum 2012; Vladimír Kyjonka; 2012.

[11] V koži pilota high-end stíhačky; ITnews; Vladimír Kyjonka; 2013; online source: http://www.itnews.sk/2013-02-04/c153968-v-kozi-pilota-high-end-stihacky

[12] What is Big Data; IBM; 2012; online source: www.ibm.com/software/data/bigdata/

[13] CRM Data Strategies: The Critical Role of Quality Customer Information; Gartner Inc.; 2003.

[14] Bitpipe research guide: Business Intelligence Overview; online source: http://www.bitpipe.com/bi/bi_overview.jsp

List of Figures

List of Tables

Attachments

The CD is enclosed as an attachment of this thesis. The content of the thesis is briefly explain in following table:

Location	Content
\Input\Cars	Sample and Metadata of input dataset Cars
\Input\ToyCorps	Sample and Metadata of input dataset ToyCorps
\Output\Visualisations\Cars	Visualisations of Analytical Scenarios on dataset Cars
\Output\Visualisations\ToyCorps	Visualisations of Analytical Scenarios on dataset ToyCorps
\Output\Reports\Cars	Reports on dataset Cars
\Output\Reports\ToyCorps	Reports on dataset ToyCorps

Appendix A

The content of this appendix is dedicated to the architecture and technologies of environment for analytical platform SAS Visual Analytics.

PROC OLIPHANT

A Proc Oliphant (see point 1 in Figure 21) is used to add and delete (data or tables) from HDFS, to add SAS data sets from any data source into an HDFS cluster, to retrieve metadata about the tables in a HDFS cluster. A table loaded into an HDFS cluster using PROC OLIPHANT has metadata information transferred to HDFS. HDFS writes are inherently slow, because all data must be written to the NameNode which in turn sends the data to the DataNodes. OLIPHANT changes this by "spraying" data to each DataNode where it is written locally then concatenated together.

sashdat – SAS Hadoop Data

A sashdat (see point 2 in Figure 21) is used to optimise to work with the LASR Analytical Server. It is created with OLIPHANT. Metadata is embedded in each block of a .sashdat file, which allows LASR to easily access the data without external file reference. It applies proper OS permissions to the file that HDFS lacks.

PROC LASR

A Proc Lasr (see point 3 in Figure 21) is a procedure used to lift an HDFS resident data set into memory creating a SAS LASR Analytic Server. The LASR starts a server instance in-memory using all machines in the cluster. PROC LASR creates a SAS LASR Analytic Server file that is accessible by the workspace server machine and the Web application server configured with SAS Visual Analytics. The SAS LASR Analytic Server file includes the host names for the machines used, the location of the Table Signature files and the port number that the server is listening on. The Table Signature files are created on the root node. Table Signature Files hold a token that enables access to the in-memory data. Authorization to read the files equals authorization to access the data.

SAS LASR Analytic Server

The SAS LASR Analytic Server (see point 4 in Figure 21) is a read-only, stateless, in-memory analytic platform that provides secure, multi-user, concurrent access to data loaded into memory in a distribute computing environment. A SAS client that runs the LASR procedure creates a file known as a SAS LASR Analytic Server file. Access to a specific LASR server requires access to the SAS LASR Analytic Server file. The server file provides information about how to connect to the root node. Table Signature files are stored on the root node. The path to the files is identified by the PATH= statement that was used when the server was started. Table Signature files are secured through operating system permissions and determines a user's ability to read a LASR Server table.

Hadoop Distributed File System (HDFS)

The NameNode is the centrepiece of the HDFS file system. It keeps the directory tree of all files in the system (see point 5 in Figure 21). It tracks where file data is kept across the cluster. It does not store the data of these files itself. DataNodes store data in the HDFS cluster. The SSH Key is generated on Head Node (also known as Root Node). The key is then distributed to each worker node to allow for secure communication from the head node out to each worker node.

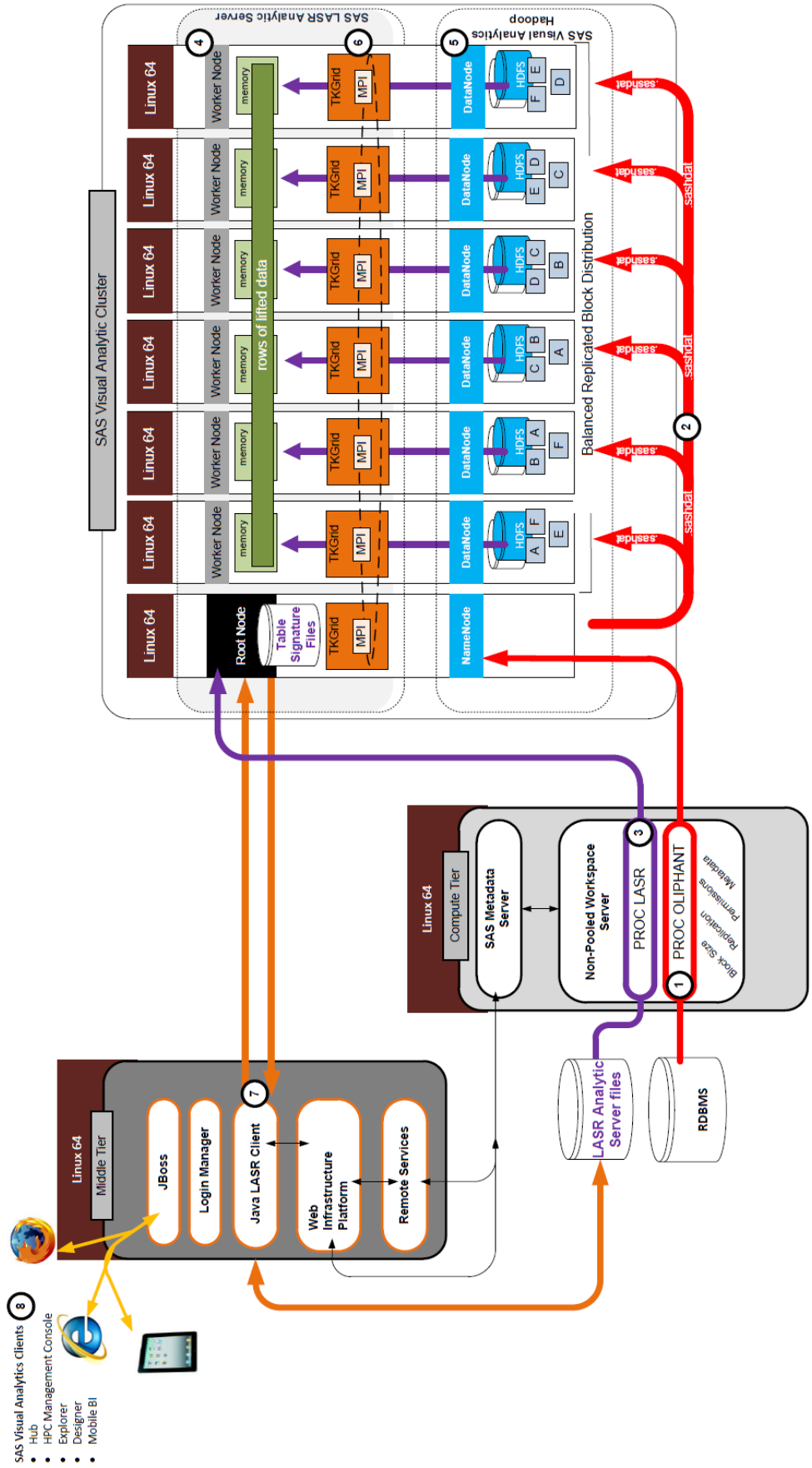


Figure 21: Architecture overview of SAS Visual Analytics.

Hadoop operating environment requires 300MB of disk space on the Hadoop partition of each node for its system files. Root user must have passwordless SSH access between all Hadoop nodes. The directory structures must be identical across all machines/blades in the HDFS cluster. NameNode Directory: "/hadoop/hadoop-name" persistent store for namespace and transaction log. Data Directory: "/hadoop/hadoop-data" where DataNodes and OLIPHANT write Hadoop blocks. Local Directory: "/hadoop/hadoop-local" where MapReduce data is written, installation requires its presence, no used by SAS software. System Directory: "/hadoop/hadoop-system" MapReduce system file location, installation requires its presence, not used by SAS software. Hadoop must be stopped when applying any updates.

TKGrid and Message Passing Interface (MPI)

TKGrid (see point 6 in Figure 21) allows for the execution of partitioned analytics to run on a cluster. MPI allows the nodes of the cluster to communicate with one another.

Java LASR Client

A Java LAST Client (see point 7 in Figure 21) is a middle-tier based Java Interface (API) to SAS Visual Analytic Server. It reads the SAS LASR Analytic Server files. It communicates to the LASR Root Node. It uses SSH to retrieve the contents of the Table Signature file.

Visual Analytic Clients

Visual Analytics works with web based client access via IE and or Firefox (see point 8 in Figure 21). The clients contain following client tools. Visual Analytics Hub (VA-H) is the central entry point for access to all role based view and functions. Visual Analytics Explorer (VA-E) is used for ad hoc data discovery and visualization to allow users to explore and analyse their data. Visual Analytics Designer (VA-D) is for the creation of reports and dashboards for the mobile client. Mobile BI is the native iPad application available from the Apple iTunes store that allows the viewing of reports created through VA-D. Data Preparation is the set of capabilities that IT will use to manage users, blacklist mobile devices, monitor servers etc. There are also data-related administration tasks to load data into the SAS LASR Analytic Server instances, perform joins and create calculated columns.

Appendix B

Dataset of the Car Sales and Marketing (Cars)

A dataset used for task A is model of automotive sales and marketing data. Model contains four main entities, such as Car Dealer, Customer, Vehicle, Campaign, and Event. Further, the Car Dealer entity contains information about location, the Customer entity contains information about personal details (name, birthday, birthday code, gender, etc.), living location (region, city, postal code, etc.), the Vehicle entity contains information about type, class, and a production date. The Campaign entity contains start date, end date, and success of the campaign, number of events within campaign. The Event entity type of event (purchase, service), financial value, and date.

The data model is de-normalised and overview of the entity attributes is given in Table 14. With de-normalisation, structure of entities with attributes is transformed into variable (each attribute becomes variable). Each variable of entity conveys information about its data type, type of variable, and unique count (output of a data profiling).

There are more two types of variable Category (Element of Dimension in OLAP terminology) and Measure (Measure in OLAP terminology). There are two common variables, which are typically being predefined and supported by exploration techniques such as time variable, or geographical variable, due to common use among different data models and special way of visualisation. Another special variable is calculated variable, the value is calculated on-the-fly rather than being stored.

Table 14: Attributes of Entities for Dataset A

Variable	Data Type	Type	Unique Count	Note
Datum kampaně	date	Category	2	
Datum konce kampaně	date	Category	2	
Datum narození	date	Category	>5000	
Datum registrace	date	Category	577	
Datum události	date	Category	789	
Datum výroby	date	Category	701	
Id dealera	numeric	Category	431	
ID kampaně	numeric	Category	2	
Id vozidla	numeric	Category	>5000	
Id zákazníka	numeric	Category	>5000	
Jméno osoby	text	Category	>5000	
Kód kraje	numeric	Category	14	
Kód kraje dealera	numeric	Category	14	
Kód obce	numeric	Category	1998	
Kód obce dealera	numeric	Category	323	
Kód okresu	numeric	Category	77	
Kód okresu dealera	numeric	Category	77	
Kód skupiny vozu	numeric	Category	11	
Kód třídy vozu	numeric	Category	6	
Měsíc události	date	Category	27	Calculated variable
Název dealera	text	Category	428	Geographical variable

Název kraje	text	Category	14	Geographical variable
Název kraje dealera	text	Category	14	Geographical variable
Název obce	text	Category	1855	Geographical variable
Název obce dealera	text	Category	322	Geographical variable
Název okresu	text	Category	77	Geographical variable
Název okresu dealera	text	Category	77	Geographical variable
Pohlaví osoby	text	Category	3	
Příjmení osoby	text	Category	>5000	
PSČ	text	Category	1053	
PSČ dealera	text	Category	301	
Rodné číslo	text	Category	>5000	
Rok události	date	Category	3	Calculated variable
Typ události	text	Category	3	
Finanční objem (tis Kč)	numeric	Measure	>5000	
počet událostí	numeric	Measure	>5000	
Souhlas k oslovení	numeric	Measure	2	
Souřadnice kraje - délka	numeric	Measure	14	Hidden item
Souřadnice kraje - šířka	numeric	Measure	14	Hidden item
Souřadnice kraje dealera-délka	numeric	Measure	14	Hidden item
Souřadnice kraje dealera - šířka	numeric	Measure	14	Hidden item
Souřadnice obce - délka	numeric	Measure	1999	Hidden item
Souřadnice obce - šířka	numeric	Measure	1999	Hidden item
Souřadnice obce dealera - délka	numeric	Measure	323	Hidden item
Souřadnice obce dealera - šířka	numeric	Measure	323	Hidden item
Souřadnice okresu - délka	numeric	Measure	77	Hidden item
Souřadnice okresu - šířka	numeric	Measure	77	Hidden item
Souřadnice okresu dealera - délka	numeric	Measure	77	Hidden item
Souřadnice okresu dealera - šířka	numeric	Measure	77	Hidden item
Spokojenost zákaníka	numeric	Measure	>5000	Calculated variable
Spokojenost zákaníka raw	numeric	Measure	101	
Událost v kampani	numeric	Measure	2	
Váha dealer	numeric	Measure	50	Auxiliary Calculated variable
Váha kraj	numeric	Measure	14	Auxiliary Calculated variable
Váha okres	numeric	Measure	15	Auxiliary Calculated variable
Zákazník kontaktován	numeric	Measure	1	

Získání/ztráta souhlasu	numeric	Measure	2
Úspěšnost kampaně	numeric	Measure	2

Dataset of Manufacturing Corporate (ToyCorps)

A dataset used for task B is a model of the corporate manufacturing and production data. Model contains three main entities, such as Product, Unit, and Facility. Further, the Product entity contains information about a material cost, a final price, and a product quality. The Unit entity contains information about its production capacity, lifespan, and reliability. The Facility entity represents a branch of corporation and a collection of Units. In particular it contains information about locations and expenses.

Again, the data model is de-normalised and overview of the entity attributes is given in Appendix B in Table 15. There are two common variables, which are typically being pre-defined and supported by exploration techniques such as time variable, or geographical variable, due to common use among different data models and special way of visualisation. Another special variable is calculated variable, the value is calculated on-the-fly rather than being stored.

Table 15: Attributes of Entities for Dataset B

Variable	Data Type	Type	Unique Count	Note
Date	date	Category	>5000	
Date by Month	date	Category	384	
Date by Year	date	Category	32	
Day of week	text	Category	7	
Facility	text	Category	21	
Facility City	text	Category	21	Geographical variable
Facility Region	text	Category	4	Geographical variable
Facility State	text	Category	13	Geographical variable
Facility Type	text	Category	1	
Product	text	Category	23	
Product Brand	text	Category	2	
Product description	text	Category	105	
Product line	text	Category	4	
Unit	text	Category	666	
Unit Status	text	Category	5	
Employees Used	numeric	Measure	>5000	
Expenses	numeric	Measure	>5000	
Expenses (capital)	numeric	Measure	819	
Expenses (material)	numeric	Measure	>5000	
Expenses (operational)	numeric	Measure	>5000	
Expenses (staffing)	numeric	Measure	>5000	
Facility Age	numeric	Measure	32	
Facility ID	numeric	Measure	>5000	De facto Category, modelled as Measure

Product ID	numeric	Measure	>5000	De facto Category, modelled as Measure
Product Material Cost	numeric	Measure	219	
Product Price (actual)	numeric	Measure	403	
Product Price (target)	numeric	Measure	186	
Product Quality	numeric	Measure	4	
Product Quality Class	numeric	Measure	4	Calculated variable
Profit	numeric	Measure	>5000	
Revenue	numeric	Measure	>5000	
Unit Age	numeric	Measure	10	
Unit Capacity	numeric	Measure	8	
Unit Downtime	numeric	Measure	2	
Unit ID	numeric	Measure	666	De facto Category, modelled as Measure
Unit Lifespan	numeric	Measure	10	
Unit Lifespan Limit	numeric	Measure	3	
Unit Reliability	numeric	Measure	240	
Unit Yield (actual)	numeric	Measure	852	
Unit Yield (rate)	numeric	Measure	>5000	
Unit Yield (target)	numeric	Measure	199	

Appendix C

The subset of outputs from experimental assignment is included in this appendix. Visualisations and reports are considered as outputs. Some outputs are already presented in Section 7.4. The completed set of output can be found in Attachments.

Drill Down

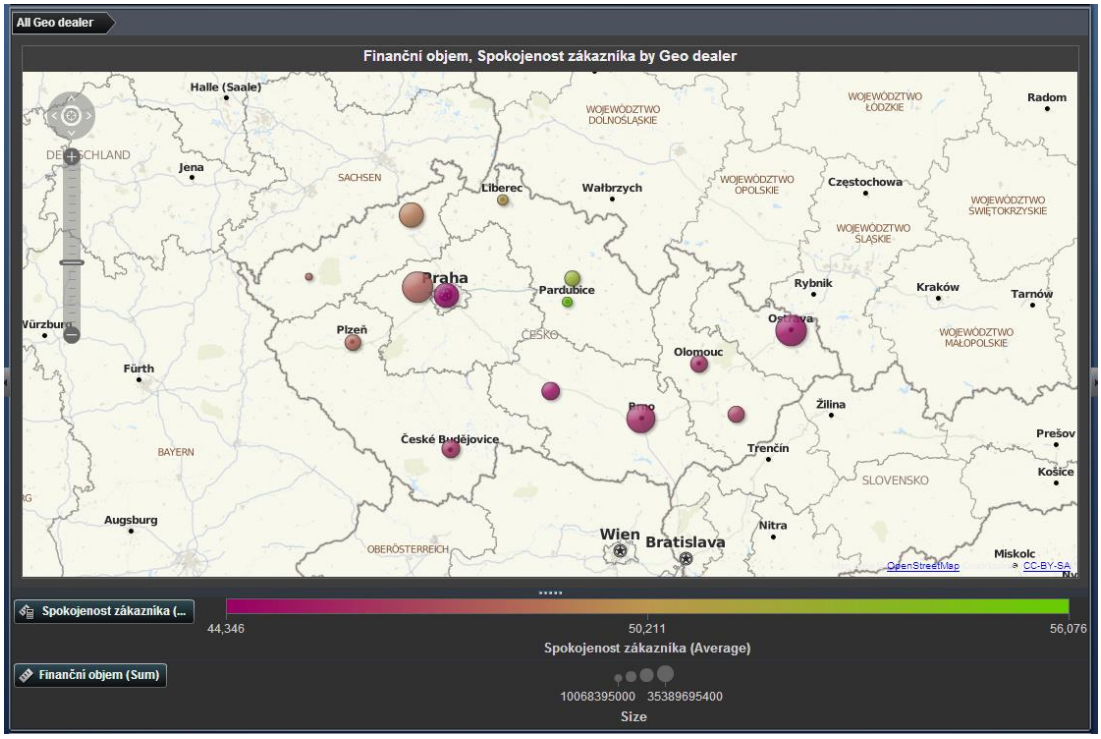


Figure 22: Sample of Geography Map.

Figure 23: Sample of Geography Map after Drill Down operation.

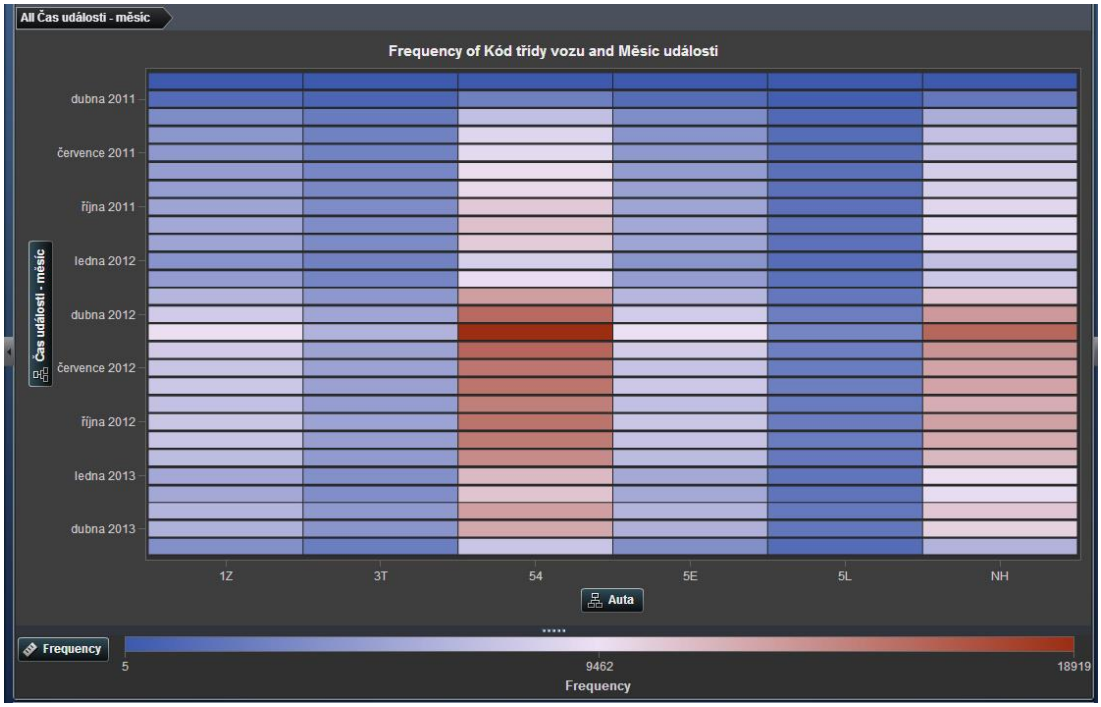


Figure 24: Example of Heat Map indicating frequency in pairs of variables.

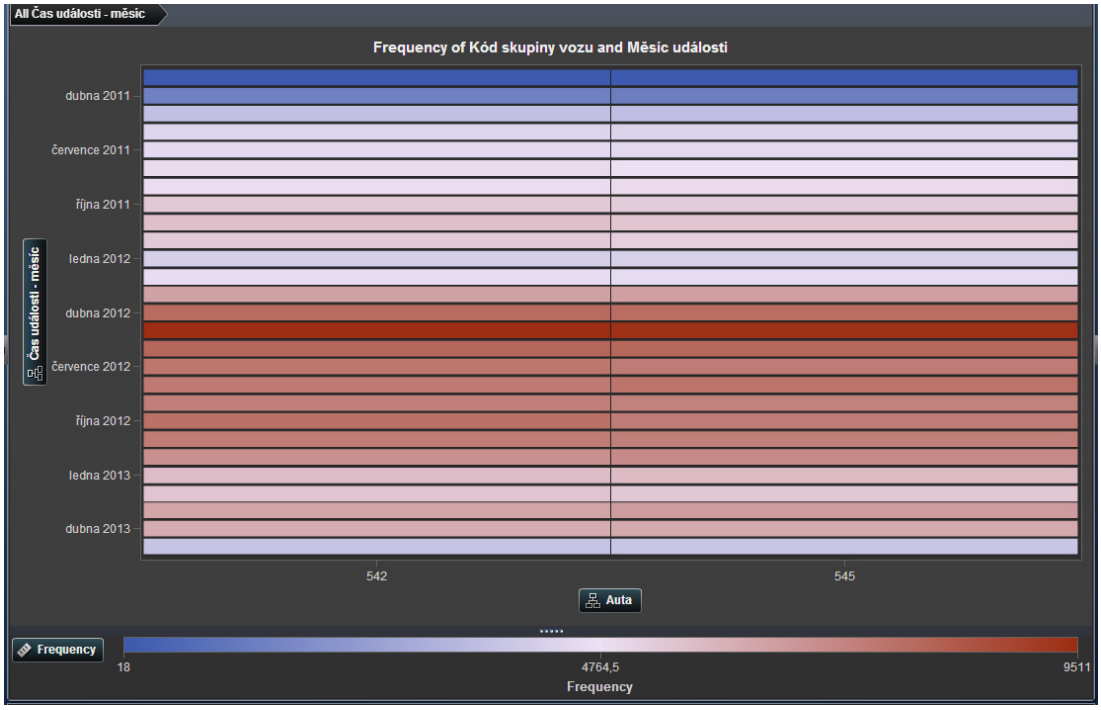


Figure 25: Example of Heat Map after Drill Down operation (from example above).

Multi-dimensions

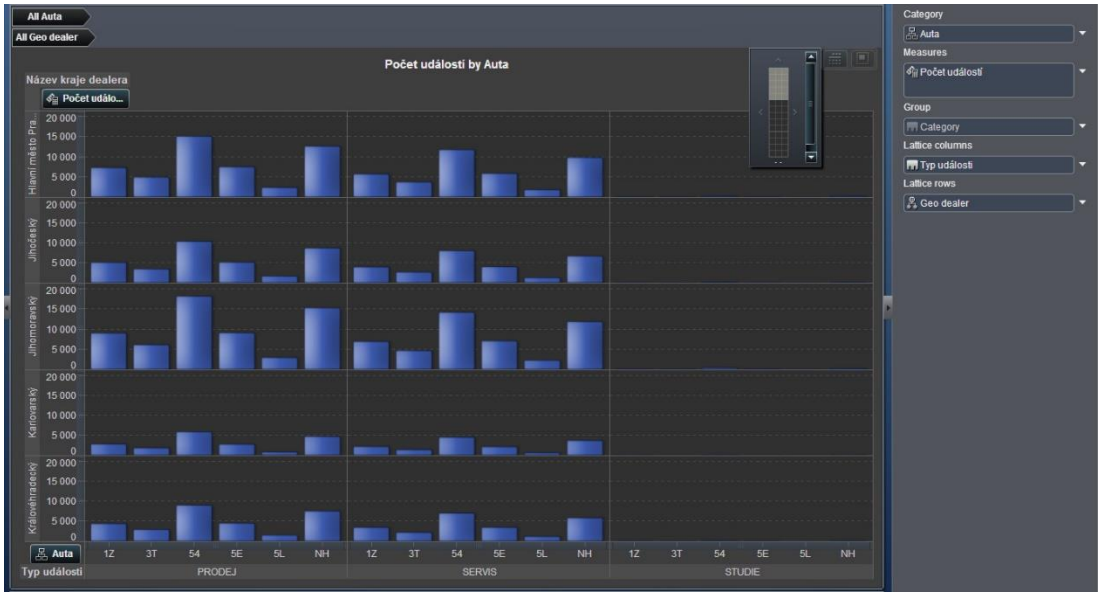


Figure 26: Bar Char Lattice with three dimensions and one measure.

Animation

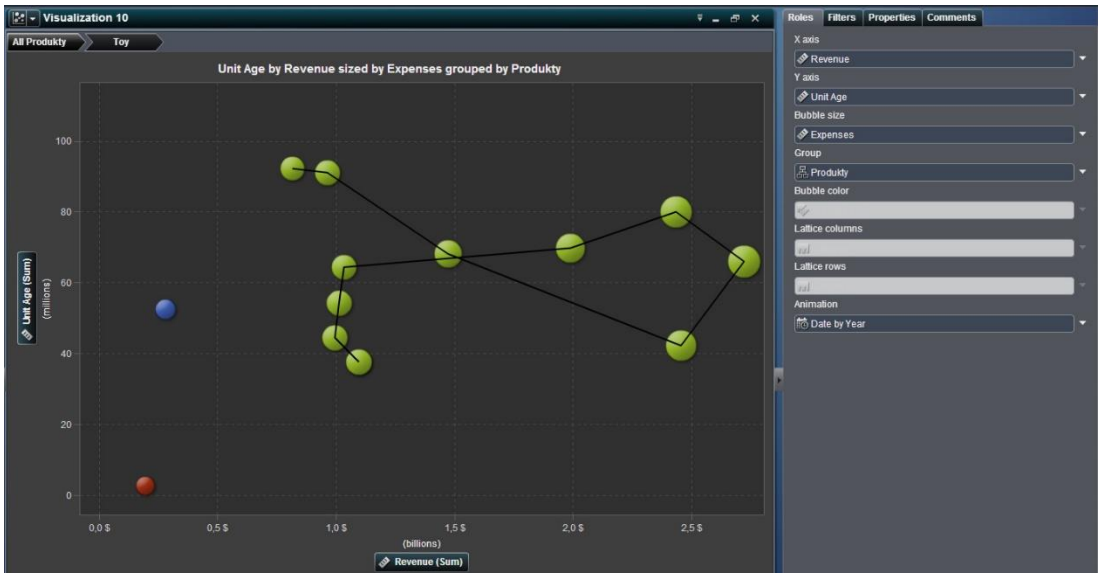


Figure 27: Bubble Plot with animation over time.

Forecast

Figure 28: Forecasting

Figure 29: Forecasting after Drill Down operation (from example above) for the same period of month.

Correlations

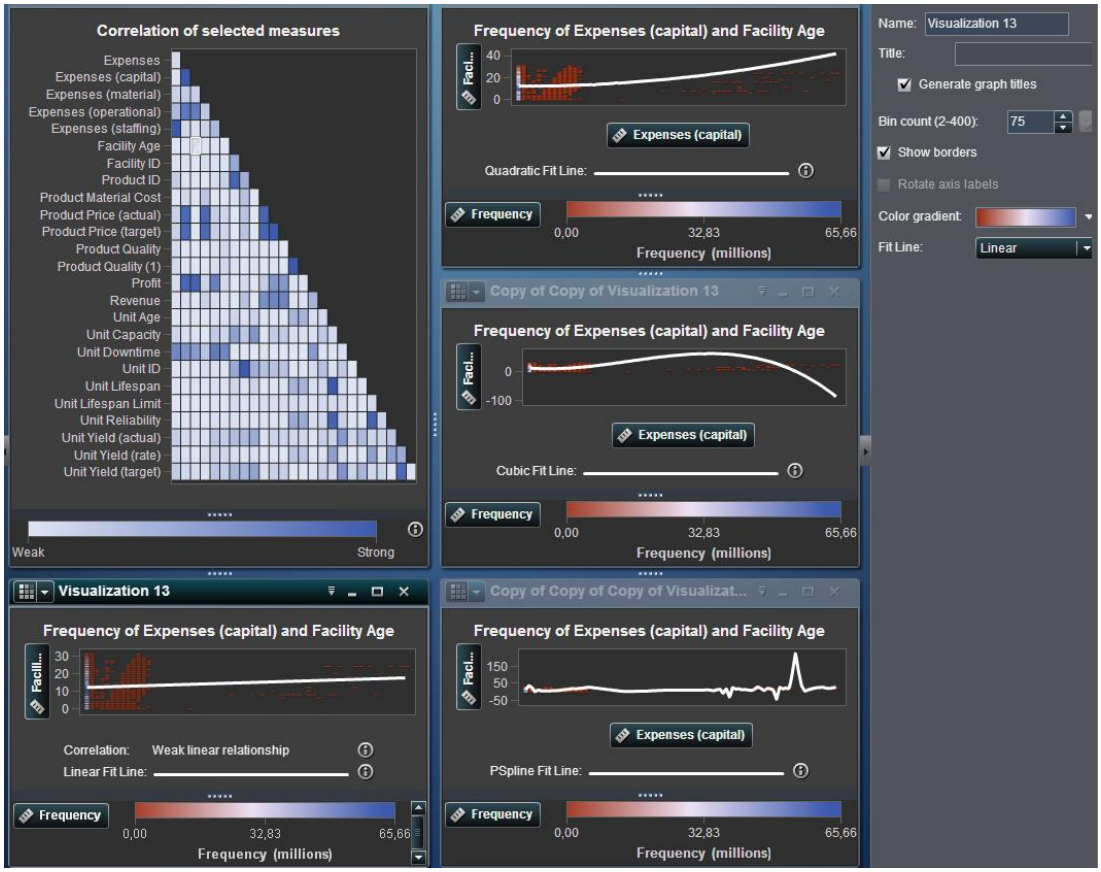


Figure 30: Correlation matrix accompanied with trend lines for specific pairs of measures.

Statistical

Figure 31: Box Plot calculating min and max value, 25th and 75th percentile, median, mean for 2 dimensions and 2 measures.

Crosstab

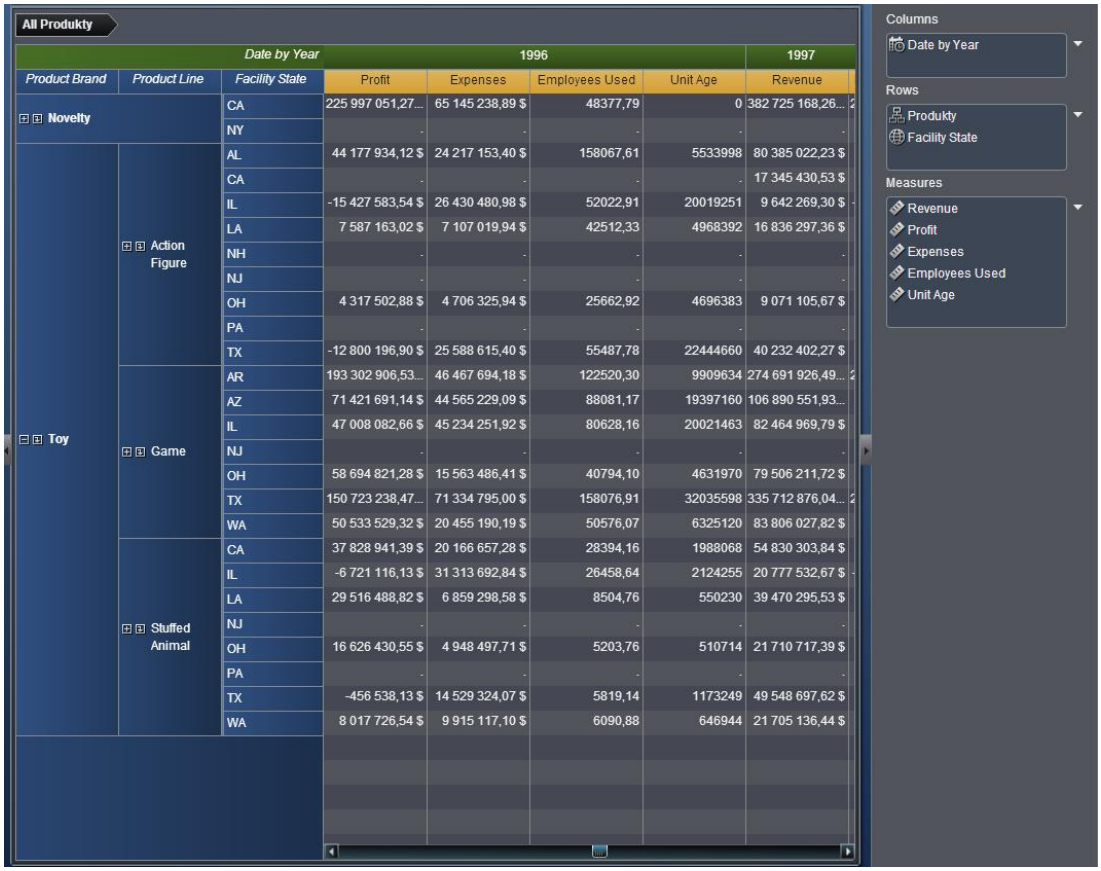


Figure 32: Example of Crosstab showing aggregated values for 3 dimensions and 5 measures.