Big Data & Data Warehouse Integration

INSY 5337 – Spring'18

Gayathri Etraj Janaki

Contents

Abstract:

Data warehouse plays a very important role in organizations in managing structured and operational data which is in turn utilized by the business analyst to analyze crucial data and its trends. Key organizational decisions are taken by the top level administration based on this data. However, due to the increase in the data available on internet and social media, also known as big data, throws a challenge upon the traditional data warehouse in analyzing data. In addition to that, it also revels shortcomings posted by software and hardware platforms that are utilized for building traditional data warehouse solutions.

In order to get a complete business intelligence advantage from big data calls for some changes in the tools and best practices for the enterprise data warehouses. For instance, in order to decide upon the useful sources for BI and EDW, BI professionals have to think outside the box. Big data has a wide range of sources, from semi-structured to structured to unstructured, and majority of the EDWs are not built to store and manage this huge amount of data. In a way to make complete utilization of big data, enterprise will have to alter their best practice.

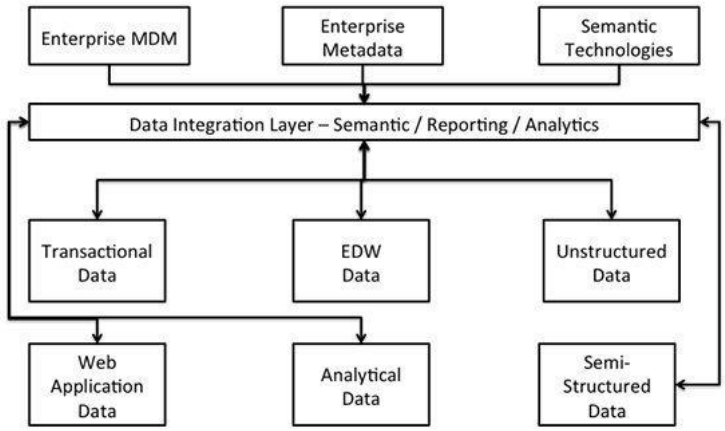
The association between big data and date warehouse is combining to become one single unit. The structured data remains in the data warehouse, whereas the divergent data is managed by Hadoop-based system.

Integrating data warehouse architecture with big data technology:

Identifying and classifying analytical processing requirements for the entire set of data elements at play is an important requirement in the design phase of the new age data warehouse platform. The support for this requirement arises from the fact that you can create analytics at the data discovery level, which is very focused, and consumer driven and is therefore not enterprise focused. Analytics can be created even after data acquisition,

Figure 10.3.2 depicts the analytics processing in the next-generation data warehouse platform. The data integration Layer is the main architecture integration layer which is a union of Semantic, Reporting and Analytics layer.

10.3.2



Once the data architecture is finalized, it will help in providing a strong base for the physical layer. The physical layer will be formed using earlier technology which may include big data as well as RDBMS systems.

Physical component integration and architecture:

The new age data warehouse will be set up on a heterogeneous infrastructure and architectures that integrate both unstructured data and big data into one scalable and high performing environment. There are several options to display the physical architecture which have advantages and disadvantages as well.

The next-generation data warehouse platform with data loading, availability, data volume, storage performance, scalability, diverse and changing query demands against the data, and the working cost of maintaining the entire data warehouse environment will be the primary challenge for the physical architecture.

Data loading:

• The loading process for big data is simply acquiring the data and storing it with no definitive format or metadata or schema. When you want to process real-time feeds into the system, while processing the data as it may be huge in number, this task can be very difficult. A gadget can be arranged and tuned to address these rigors within the setup, as contradicted to a pure-play execution. The flipside is that a custom design configuration may happen, but this could be illuminated effectively.

• In the case of large files, continual processing of data in the platform can create contention for resources over a period. If this is a key prerequisite, a device can be suitable for this specificity, as the guessing game can be evaded fully.

• MapReduce configuration and optimization can be difficult in high-computing surroundings the appliance architecture provides you reference architecture setups to avoid this danger.

Data availability:

• Data accessibility could be a challenge for any framework that relates to preparing and changing data for utilize by conclusion clients, and big data takes after this standard. The good thing about Hadoop or NoSQL is to diminish this chance and make data accessible for examination promptly upon acquisition of information. The challenge is to stack the information rapidly as there's no pre-transformation required in this use-case.

• Accessibility of data will for the most part depend on the quality of metadata within the SerDe or Avro layers. On the off chance that data can be enough changed on procurement, it can be accessible for examination and revelation quickly upon securing of the data.

• Since there's no update of data within the big data layers, converting new data ordinarily contains updates which is able make copy information, and thus this has to be dealt with carefully to decrease the affect which may exist due to non-conversion.

Data volumes:

• Big data volumes can effectively get out of control due to the natural nature of it. Cautious consideration ought to be paid to the development of data upon each cycle of securing of data.

• Conservation necessities for the data can change which may depend on the nature of the data and the recency and its importance to the commerce system.

Compliance requirements:

Data security and storage can be directly affected by Safe Harbor, SOX, HIPAA, GLBA and PCI guidelines. If you are preparing to use these data types, one must design accordingly.

Legal mandates:

There are different value-based information sets that were not put away online and were required by courts of law for disclosure purposes in class-action claims. The big data framework can be utilized as the storage engine for this specific data type, but the data needs certain compliance needs and extra security as well for it. This information volume can affect the generally execution, and on the off chance that such information sets are being prepared on the enormous information stage, the apparatus arrangement can give the directors with instruments and tips to zone the foundation to stamp the data in its possess zone, diminishing the chance as well as the execution.

• Data investigation and mining is the foremost common action which aids huge organizations to obtain information conjointly produces expansive data sets as the output of it. These data sets ought to be maintained within the big data framework by intermittently cleaning and erasing intermediate data sets. This is often a range that ordinarily is overlooked by organizations and can be a noteworthy deplete on the execution over a period of time.

Storage performance:

• Disk execution is an imperative figure to considered and the apparatus demonstrate can give a much better center on the capacity lesson and multi- tiering design. Long-term arranging and development administration of the capacity foundation can be accomplished by this.

On the off chance that a combination of in-memory, SSD and conventional capacity architecture is arranged for big data preparing, the perseverance and trade of data over the diverse layers can be time-devouring both handling time and cycles. Care must be utilized particularly in this region, and the apparatus design gives a premise for such tall capacity necessities.

ARCHITECTURE

From the point of view of the logical abstraction of architecture, both the Data Warehouse and Big Data have the same components: Data Source, Extraction, Transformation and Loading processes (ETL), storage, processing and analysis. Due to this, an overview of the architecture in terms of this components are presented below.

Sources and data types:

Fourth Generation Technologies asssited the growth of transactional applications that allowed the automation algorithms on repetitive structured data. Structured data (SD) is distinguished for being well defined, predictable and soundly handled by a difficult infrastructure.

Technological developments, digitization, hyperconnected devices, and social networks, among other technology enablers brought unstructured information to the scope of enterprises. This includes information in digital documents, data coming from autonomous devices (sensors, cameras, scanners, etc.), and semi-structured data from web sites, social media, emails, etc. Unstructured data (USD) doesn't have an unsurprising and computer recognizable structure and may be divided into repetitive and non-repetitive data. Unstructured repetitive data (US-RD) are data that occur in many occasions in time, may have a similar structure, are generally massive, and not always have a value for analysis.

Samples or portions of these data can be used. Due to its repetitive nature, processing algorithms can be doubted for repetition and reutilization. A typical example of this category is data from sensors, where the objective is the analysis of the signal and for which specific algorithms are defined. Unstructured unrepetitive data (US-URD) have varying data structures, which implies that the algorithms are not reusable (and the task of predicting or describing its structure is already a complex one). From our perspective, besides free-form text, imagery, video and audio also pertain to this category. Traditional Data Warehouses were born with the purpose of integrating structured data coming from transactional sources and to count with historical information that is supported by OLAP-based analysis.

Extract-Transform-Load processes (ETL):

Construction of Data Warehouse requires Extraction, Transformation and Loading processes (ETL). These must consider various other factors such as several data quality related issues, as for

instance duplicated data, possible data inconsistency, huge risk in data quality, garbage data, creation of new variables using transformations, etc. This increases the need of specific processes to extract enough and necessary information from the sources and implementing processes for cleansing, transformation, aggregation, classification and estimation tasks. All these, despite the utilization of different tools for the various ETL processes, can result in fragmented metadata, inconsistent results, rigid models of relational or multidimensional data, and thus lack of flexibility to perform generic analysis and changes.

Thus, the need of more adjustable ETL processes and increased performance gave birth to proposals such as real time loading instead of batch loading. Middleware, for instance the engine of flow analysis, was also introduced. This engine makes a in-depth exploration of incoming data (identifies atypical patterns and outliers) before it can be integrated into the cellar. On the same line is the Operational Data Storage (ODS), that proposes a volatile temporal storage to integrate data from different sources before storing it in the cellar. The work presented in various traditional architectures, creates an ETL subsystem in real time and a periodic ETL process. Periodic ETL refers to the periodic importation in batch from the data sources and the ETL in real time. Using Change Data Capture (CDC) tools, changes in the data sources are automatically detected and loaded inside the area in real time. certain conditions are met, data are loaded in batch into the cellar. The stored part can be then divided in real time and static area.

Concentrated queries for sophisticated analyses are made about storage in real time analysis. Static data is equal to Data Warehouse and historical queries are thus handled in the traditional way. It`s worth to mention that for a Data Warehouse few changes had been observed which may include temporal processing areas and individualized processes.

ETL requirements for new data:

With the need to manage unstructured repetitive data (US-RD) and unstructured unrepetitive data (US-URD) coming from different sources (like the previously mentioned) new requirements are bound to rise which may include the following:

- Managing rampant data growth. Data Warehouse is used for using transactional databases of the organization as the main source, eventually flat files, and legacy systems. While data volume increases, it increases at a manageable pace. The new Data Warehouse and Big Data also provide solutions to the management of large data collection by the MapReduce programming model, which may allow to parallelize the process for then gather the partial results; all this supported in a distributed file system like Hadoop Distributed File System (HDFS).
- The frequency of arrival. This can range from frequent updates to bursts of information. Traditional Data Warehouse normally does not face this problem, since it always focuses on data that can be extracted or loaded in a periodic and programmed way. Since Big Data was designed to receive all the incoming information at any moment in time, it must use any required amount of memory, storage and processing.
- Longevity, frequency and opportunity of use. Statistically, the most recent generated dataset will be used more frequently and in real time. Once datasets become old, their frequency also decreases. However old data can always be used for historical analysis.
- Integration of data. While the traditional DW was intended to integrate data across a multidimensional model, the appearance of unstructured repeated data (US-RD) raise problems which are related to find necessary ways to group the data under a context which is independent of the data type. For example, to group pictures with dialogues, even within the same type of data (to determine the context of an image by the same image) to find the structure that best represents all the data and do algorithms to integrate, transform and represent such data.

- For Disorganized unrepetitive data, apart from identifying the context and structure, an algorithm for each dataset may also be required, which prevent the reuse and it also increases the difficulty. The integration of diverse datasets is the significant difference between a Data Warehouse and Big Data. In Data Warehouse, the underlying purpose of integration is to have a unified vision of the organization, while in Big Data, integration is not the goal. For Big Data, some unstructured datasets not amenable to integration should be kept in raw format, allowing the possibility of further uses that may be not foreseeable now.

ETL for new data life cycle:

The life cycle of data life cycle comprises of three stages

First is the interactive sector where most of the new data is present, the update is performed online and a high response time in performance is required. Secondly, the integration sector where the interactive sector data is integrated and transformed, and data is present which may not depend on the need of an organization. Third the archive sector, which maintains historical data and has a lower access of probability. The data highway, consisting of 5 caches which are given below, and it is sequentially arranged as per the frequency and longevity of data.

a) raw and immediate use of data;

b) Data and frequency of usage (in seconds);

c) Data Used for business monitoring and frequency of usage in minutes;

d) Data Used to generate reports for business decisions and frequency of usage:

e) Aggregated Data for Historical Analysis.

A reference architecture based in components that allow handling of all kinds of data: acquisition, cleaning, integration, identification, analysis and management of data quality. It also includes traverse components for data storage, metadata as well as the lifecycle and security handling of data. As for products which are already in the market, it`s worth to mention that Oracle implements a global proposal including both ordered and unordered data as well as different storage areas where the lifecycle of the data is present. The proposal relies in a set of tools that may permit data gathering, organization, analysis and visualization.

SAP Data Warehousing offers a different solution that integrates features of Big Data and Data Warehouse in real time allowing the analysis and identification of patterns in complex structures of both structured and unstructured data. SAP supports ETL processes in the SAP NetWeaver tool, which allows to integrate data from different sources. Pentaho as free software platform includes the component that allows to do the typical ETL processes for Data Warehouse and, through Hadoop, supports ETL for Big Data. The above solutions consider the life cycle of the data which are technology focused.

Storing, processing and analysis:

Data Warehouse systems have traditionally been supported by predefined multidimensional models (star and snowflake) to support Business Intelligence (BI) and decision making. These models are generally used on relational databases and accessed through SQL. Less frequently, implementations under multidimensional schemes (Multidimensional Online Analytical Processing MOLAP) are also found. Although the traditional Data Warehouse manage huge data, the architecture is supported on client-server models which can only be scaled in a vertical way, which implies huge technological and economic efforts which may include the development and the maintenance.

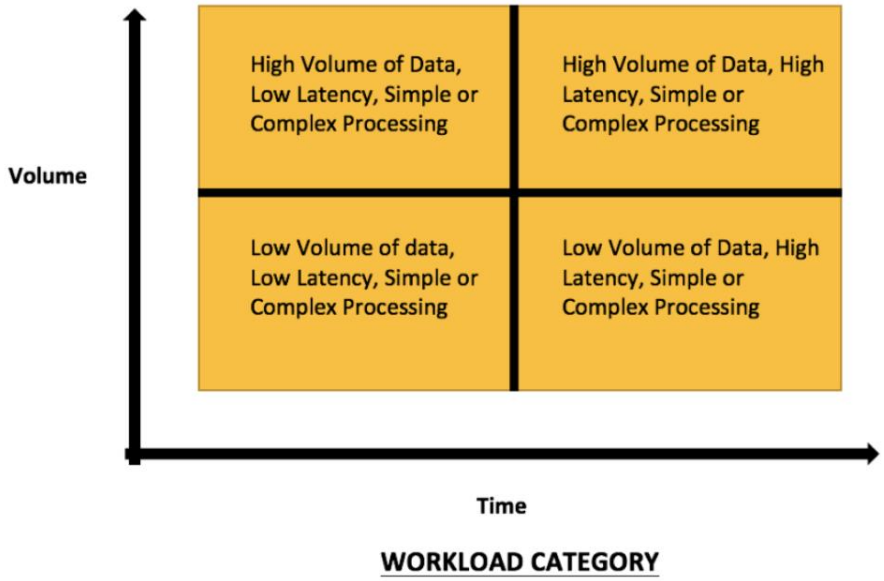
Contrary to this, Big Data and the new Data Warehouse generation neither have preexisting analytical models nor do they rely on client-server architectures and they must support the horizontal scaling. The answer to the new needs is the use of exhaustive memory, data distribution and processing parallelization, which in one way or another are included in Hadoop, MapReduce, NoSQL databases, Storage repository and processing in memory and technologies help in achieving the objective.

INTEGRATION STRATEGIES

An architecture that is Data-driven involves analysis and design of data processing and physical architecture that merges structured OLTP data and the big data in one single domain.

Architecture

From the various data types of data components, we can differentiate between the power and an ordinary user of the data warehouse system.



Workload

While processing Big Data one of the biggest requirement is workload management.

We can execute the workload request of the types of the data because of the data architecture and categories which helps us to allocate the suitable framework. There is a total of four important categories of workload rooted on volume of data and related latencies. The type of data depending on the category can be split into physical framework layers for future operation. This method used for management makes a dynamic scalability necessity for all parts of the DW, which can be defined by curbing the current and latest infrastructure options. Important note to consider here is that the logical needs should be flexible in order to be applied over the various framework elements since the identical data might be categorized into many workloads depending on the priority of processing.

Let's take an example, processing low-latency data with less-volume, high-volume, high-latency data makes a distributed weight on the environment where processing of data takes place, which in fact, would have processed one kind of data and its workload. In addition to this, data loading happening simultaneously or in a quick interval, and now the situation can slip out of hand in a quick series and has an impact on the performance. This problem magnifies if an identical framework in employed for big data integrated Data Warehouse.

The aim for employing the workload quadrant is to identify the challenges and problems related with the processing and how to minimize the related trouble in framework design to create the future generation DW.

How can Big Data co-exist with EDW?

Data Staging:

In the average DW today, a lot of heavy data processing is done in and around data staging areas and thus data staging is not just staging data. Data Staging areas were considered temporary storing bins where ETL or data quality jobs would place data before another job (reporting, analytics) uses it for further processing. That is why data staging areas are known as "analytic sandboxes" which store data for a period before deleting. User adoption requires detailed source data and a large volume of it and hence developing big data makes this possible. Staged data is transient in nature, similar is the case with big data.

The 'T' in the ETL in data warehouse can be transferred to Hadoop. The detailed source data handled by data staging platforms is mostly structured but evolving to include more unstructured and semi-structured data. A staging platform which handle the full range of multi-structured data requires more space to hold more data for longer duration. Hence the staging area can be built on the Hadoop Distributed File System (HDFS). HDFS proves to be cost-effective, live archive for many data types.

By moving the "T" in ETL to Hadoop, organizations will be able to reduce costs and empty database capacity and resources for faster user query performance.

Schema Flexibility:

RDBMS –

Relational Database Management which is usually used in Data Warehouse Implementations are well equipped in storing highly structured data from: ERP, CRM and other operational databases. The data are mostly in semi-structured data which are in XML and JSON. Hadoop can easily and quickly ingest any data formats which has evolving schemas like A/B & multi-variate tests on many websites. And also there would be no audio, video or image schemas.

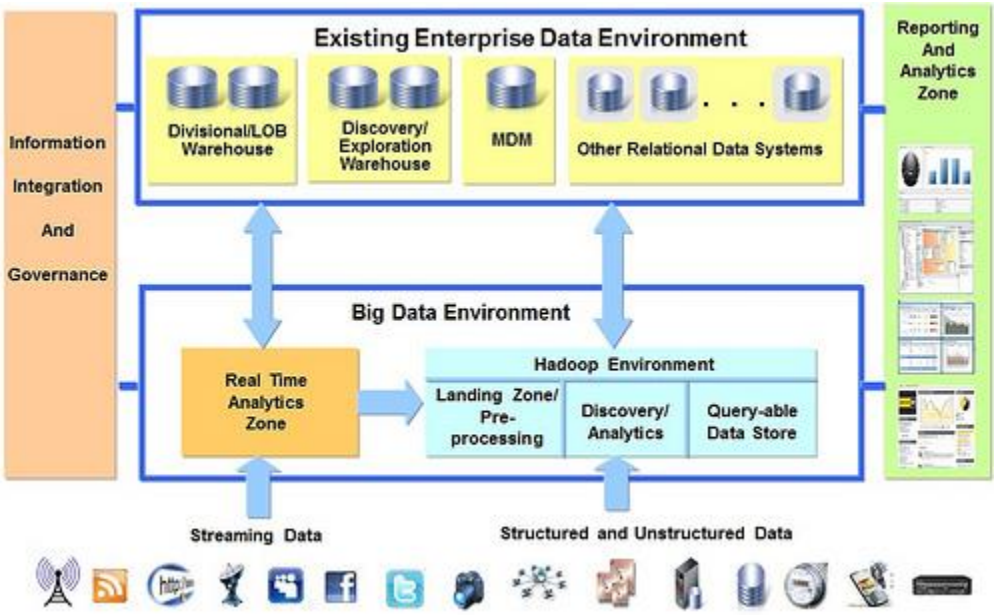
Processing flexibility:

Hadoop provides a variety of programming languages thus it comparatively has more capabilities than SQL. Hadoop's NoSQL is more of a natural framework one which controls the non-traditional data types and thus empowers the procedural handling and handling of important use-cases. It could be used in cases of time-series and gap recognition. The EDW in a firm with a Hadoop system can be empowered and improved/augment its capabilities is by following the below steps:

a) The regular OLTP structured data and back-office framework are continuously stored in the Enterprise Data Warehouse(EDW).
b) The unstructured data that does not fit the tables can also be stored in the Hadoop/NoSQL system. These unstructured data have all the data related to the communication with the

customers, their feedbacks, log details, images, emails, texts, tweets and statuses. All of this being stored into the Hadoop system.

c) The data's present in both EDW system and the Hadoop system are co-related and analysis is done. Many schemes like ad-hoc analysis, clustering & targeting of models are used against this analysis which helps us to better understand the clients, other competitive products and the results have also computationally intensive results.



This big data integration with data warehouse can be explained using two use-case scenarios:

1) Hadoop is used by a brokerage firm to reprocess its raw data. These are followed by the processing of the raw-click streams of the customers using their website. Processing of these provides very valuable insights on customer's preferences and these are sent to the data warehouse to be stored and help evaluate the changes in future. The DW not only couples the marketing campaigns according to the customer preferences but also provides valuable suggestions related to investments and also these are analyzed to provide an overview & are shared with the users.
2) Machine learning is performed using the Hadoop systems for e-commerce website services. It helps detect fraudulent suppliers. These can be extended to find also the fraudulent websites by using the Hadoop's predictive model of similar kind. This model can be loaded to the Data Warehouse where it can be used to evaluate the sales activities of the fraudulent websites. Once found they are investigated and removed.
3) The Evolution of the Enterprise Data Warehouse, starring Hadoop: http://data-informed.com/the-evolution-of-the-enterprise-data-warehouse-starring-hadoop/
4) Use case from Dell big-data solutions - Find out how to use Hadoop technology and an enterprise data warehouse in performing Web page analytics: http://www.dell.com/learn/ba/en/bapad1/videos~en/documents~hadoop-edw-web-page-analytics.aspx?c=ba&l=en&s=biz&cs=bapad1&delphi:gr=true
5) Use cases from Cloudera - Ten Common Hadoop able Problems (Real-World Hadoop Use Cases): http://blog.cloudera.com/wp-content/uploads/2011/03/ten_common_hadoopable_problems_final.pdf

The Hadoop uses the Data Warehouse as a data source and also uses to leverage the capabilities of using these two systems separately. It can be used by many companies to achieve much more potential capabilities and competitive advantages.

CHALLENGES AND WAYS TO OVERCOME:

Challenges in Architectural Integration:

In the new-age data warehouse platform, the key engineering mix layer here is the information joining layer, which is a blend of semantic, announcing, and investigative advances, which depends on the semantic learning structure, which is the foundation of cutting edge insights and business insight. Concluding that the information design is the most tedious assignment that, once finished, will give a solid establishment to the physical usage, the physical execution will be refined which may utilize advances from the prior exchanges, including Big Data and RDBMS frameworks.

The cutting-edge data warehouse center will be set on a heterogeneous foundation and designs that incorporate both conventional organized information and huge information into one versatile and multi-performing condition. Usually, associations know the arrangement of abilities they wish to convey, and they can verbalize a conclusion to-end guide. They can differentiate the various stages and assets which can be expected to fulfil the targets.

With Big Data, associations may have a thought or intrigue, yet they do not really recognize what will leave it. It requires a one of a kind blend of ranges of abilities, any semblance of which are new and not in wealth. The design improvement process should be liquid and altogether different from SDLC-like engineering process such many associations utilize today. It must enable associations to persistently survey advance, rectify course where required, adjust cost, and pick up acknowledgment.

The six key strides in the process delineated here are to build up business setting and degree, set up design vision, survey the present state, evaluate the future state model, characterize a vital guide, and set up administration over the engineering. This tends to be a shut circle process, since effective arrangement increases new thoughts for achieving business objectives. The architecture makes a aggressive intend to develop toward the future state engineering.

Key standards of the guide incorporate specialized and non-specialized developments intended to convey business esteem and at last meet the first business desires. The guide ought to contain: An arrangement of compositional holes that exist between the present and future express, a money saving advantage examination to close the holes, the incentive from each period of the guide and proposals on the most proficient method to expand esteem while limiting danger and cost, consideration of innovation conditions crosswise over stages, flexibility to adjust to new business needs and to evolving innovation. An arrangement to take out any aptitudes holes that may exist when moving to the future state (e.g. preparing, employing, and so on.)

Challenges faced in Costs:

While integrating data warehouse with big data the cost arises because of the items that are difficult to quantify and eventually to forecast. These are:

1) The labor cost which is for programming, evaluation, for the learning done in the beginning and the data which is additionally acquainted.

2) Additional software and hardware items that were procured

3) Last Minute Technological changes

4) Increase in direct cost

We have tried many things to assign the process of data maintenance but still a huge portion of it we could not be anticipate. Unrealistic expectation often results in an optimistic budget. If we fall short of funds than it could result in increase of timeliness for the completion of data warehouse. There are many things which are unanticipated those could also add to the costs. We should be pretty much accurate with our estimate with respect to extraction, maintenance and cleaning. But these estimates could differ a lot if the data given to us is not strong which may increase the integration costs.

Strategies:

1) The estimate we do for project should be far sighted and rational. To achieve this, we need analysts who are experienced the lessons learned by the companies in the industry.
2) There will be a need to identify items that we know could affect the total project cost but could not be traced in first sight while estimation of the project cost.
3) Departments should have a good interconnection and trust within them. After having a good planning, expertise opinion and having trust of the stakeholders, we should be having good management of the processes so that cost overruns could be as less as possible and if it all happens it could be overcome easily. In the processes like data integration, the cost overruns could be quite frequent so experience along with vigilant oversight is needed to as to add to the profit instead of adding the cost which may increase the budget as well.

Data Integration should work good for the users if it does then automatically it will be a fundamental to business process. It is a wrong belief that the best integrated systems are designed well. However, the best integrated system will be having a routine care and it will be supported by some experienced personnel

Data Storage Challenge:

Large amount of data is generated and used by organizations for data analytical processes. Storing such large data has become the most important challenge in Big data. For any ideal storage system, capacity, performance, throughput, cost, and scalability become the important parameters. Reliability forms another important parameter for big data. These large amounts of data are outside the limits of hardware and software. This is the reason there need to be made changes in the architecture so that we could process large amount of data and the same could be reconstructed reliably in the system.

Storage management uses technologies which help organization to improve data storage performance.

Methods to overcome:

1. **Big Data Storage Mediums:**

   The quality of the storage devices has a huge impact on the entire storage system. Access time, data transfer rate, and cost-effectiveness play an important role in the big data environment and these can be benefitted with the use of hard disk drives(HDD) and solid-state drives(SSD).

2. **Backup Strategies:**

   Recovery of data from a backup is the main objective. For a proper recovery of data, the time and method of backup are important. A timely full backup thus proves to be beneficial from the big data perspective.

Also, data mining and data exploration are two main operators of big data which produce large amounts of data. On time maintaining and cleaning of medial datasets influences the performance and storage of the system.

Data Transformation:

It's a fact that in EDW staging area, enormous amounts of data are processed to prepare and report source data for report generations, analysis of data and more. These result data are loaded into various databases such as EDWs and DataMart databases. Homegrown ETL tools are generally used for this processing. Hadoop is preferably used due its capability to be advantages compared to the traditional ETL environment, those being – its ability to be easily scalable and economical. Due to these reasons Hadoop has started to replace the ETL environment in many organizations. A good instance to support this is to dump heavy transformation, the Transformation part in the ETL, taken from the Data Warehouse into Hadoop. The organizations for years have faced many difficulties to scale the traditional ETL architectures.

The transformations are pushed to the data warehouse platforms by the data integration plans. Due to these reasons we can say assuredly that this integration in the EDW architectures would use more than 80% of the capacity of the database resources and capacity. But this situation results in raising of the costs, support services to users but provides very poor user query performance. The organizations can release some database capacity & resources and reduce costs. This would help improve user query performances.

The following steps can be used to enhance and augment EDW in an organization with big data cluster or a Hadoop system:

- The usual methods of storing structured data from OLTP and the back office systems into the EDW are done as usual.
- It does not follow the traditional structure of storing structured data in the EDW tables. Unstructured data which are customer related data from logs related to customer feedbacks and phone details, emails, images, tweets and text messages are stored into the Hadoop or NoSQL system.
- The data in the EDW is co-related and combined with the data present in the Hadoop clusters to get insights about all sorts of data about customers, products, equipment's and more. Ad-hoc analytics, clustering and targeting models are done by the organizations. These results are then compared with the data already present in the Hadoop systems. This is done because the previous data would be very much computational data and would be very intensive.

Though it is a general ideology that data transformation would be a challenge during the integration phase. But the execution of use case scenarios represents something else. EDW integrated with Big Data technology augments its benefits and provides to be the most beneficial. These two systems can work together in parallel systems and can complement each other.

EXAMPLES:

Hadoop Powered EDW optimization Solution by Hortonworks

Hortonworks vision for solutions:

Hortonworks leans completely towards supporting the open source group and is acknowledge for its Hadoop advancement in the ASF. Not long ago, Hortonworks have expanded their domain into multi-item portfolio, starting with appearance of Hortonworks Data Flow. Hortonworks recognized the importance of cloud compared to the get-go and transferred quickly with their cloud vending. Hortonworks and Microsoft have teamed up to work on Azure HDInsight which is managed by HDP for over a period of time. They recently divulged Hortonworks Data Cloud for AWS.

For technology organizations that work towards enhancing and building the alliance with clients and customers, the standard progress is to speed up the delivery of a long repeatable pattern of use. This

is the right spot for Hortonworks arrangement methodology to step in. The aim is to gather arrangement that will transmit quickened value to the clients, allowing things to be possible with the data, thus resulting spectacular business transformations over each vertical.

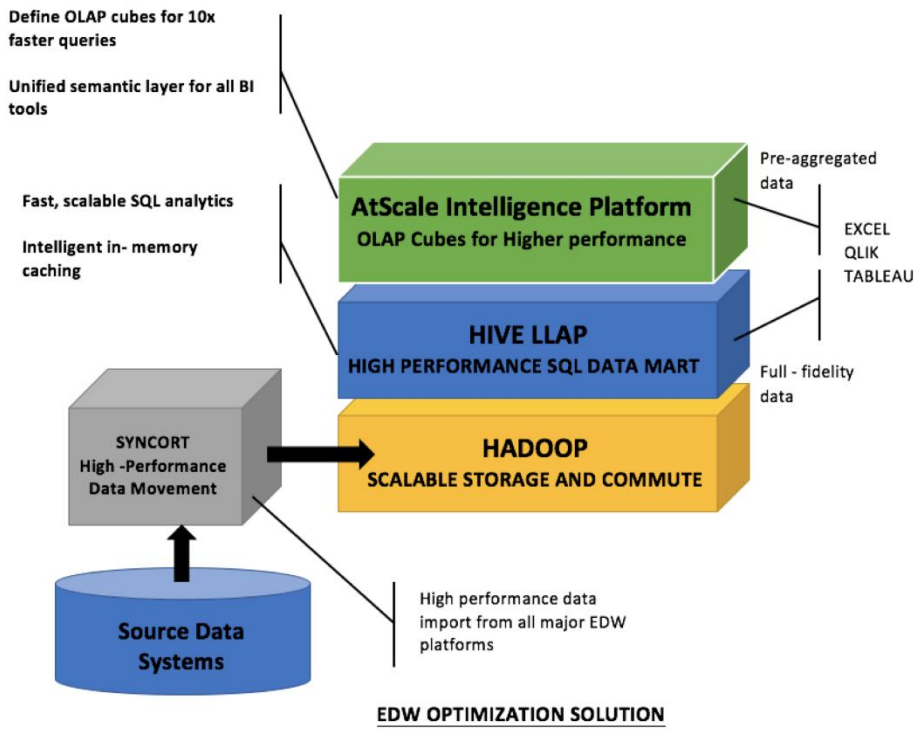
Hortonworks EDW optimization solution:

One of the most occurring usage pattern from the client is the progress of the organization towards expand their EDW surroundings, extending the life of present ventures while, at the same time delivering the business interest for agile analytics on this ever-growing data quantity.

The benefit of EDW Optimization Solution is that it allows us to store documented EDW data in Hadoop at much reasonable cost. In addition to that, it allows us to offload non-critical processing loads into Hadoop where it is significantly less costly. All in all, we have the capacity to add new information sources to Hadoop, much faster when compared to pixel perfect demonstrated EDW. **The cherry on top is to bring quick BI accessible then on top of this fine grained and enhanced information.**

- **Bring down cost storage for EDW Data**
- **Bring down cost preparing of non-critical loads**
- **Effortlessly improve EDW data with new information sources pixel something??**

Quick BI of full data set:



EDW OPTIMIZATION SOLUTION

Benefits of Hortonworks EDW optimization solution:

Empowering big data analysis is the main advantage of incorporating Hadoop into BI/DW: Generally, people perceive Hadoop's essential part as a major data source for analytics. Likewise, HDFS' versatility additionally empowers greater and better practices in data archiving and schema-free data staging even with machine data from robots, sensors, meters, et cetera.

Hadoop is useful for exploratory data analytic techniques: Enormous data and analysis go together because scientific strategies help client associations get an incentive from huge data (which is generally a cost focus) in the type of increasingly various and exact business bits of knowledge. Be that as it may, these are not the same bits of knowledge that a conventional BI/DW arrangement gives in view of precisely arranged data about well-understood business elements. Since most enormous data originates from new sources sometimes tapped by BI/DW, clients can direct data analysis with enormous data to find new certainties about a business, and in addition its clients and operations.

Hadoop scales cost successfully: HDFS is known to accomplish extraordinary versatility on low-cost equipment and programming, so it can help clients catch a bigger number of data than before.

Hadoop is useful in extending DW environments:

Numerous clients felt that Hadoop is good for EDW, because of Hadoop's emphasis on cutting-edge analytics supplements the attention on revealing and OLAP ordinary of generally DWs.

HCL EDW Optimization Solution:

The old information which are stored in the social warehouses, which are indefinitely costly, are causing a deep tackling situation with urgency in cost. This is because there is a huge volume of information that are available to be stored.

From a very long time the data storage management in all organizations has been done by the concept of data warehouses. The expenditure of the data warehouse systems can be brought down by making it legit by the organization which uses it and releasing an open stage for the research or investigation of the large information.

The concept has reached many forms and types of its own creation and level of maintenance; however, the bottom line is retained at the same level and the core concept of how data is handled are shared by all these forms. The Customer management system(CRM) is one of the great examples on how a system utilizes the business intelligence on the huge data sets and reaching a conclusion for the queries and subjects in research. This is all easier to be handled when the whole pattern of data storage and optimization is thoroughly structured and optimized. The highly centralized setups help in the whole process with a dependence on the data warehouse scenarios.

The solution in hand, The HCL Data Warehouse Optimization Solution analyzes your current data warehouse environment be it Exadata, Vertica, DB2, Teradata, Netezza, or SQL Server and offloads the fitting capacities and information to a savvy, big data platform. The data warehouse creates a humongous information stage by combining the information in a realistic way and building new progressed assessment on the collected information.

The HCL tool targets to analyze the data, identifies the possible victories from the data collected through the statistics. A roadmap is formed to cover both the checkpoints, analyzing and offloading the data. The data is migrated to the destination and it performs ETL jobs and moves the whole interface into an extension that thrives in establishing a big data COE. The establishment of data solutions forms the basis of this HCL tool. The HCL tool also adds additional data sources to this extension.

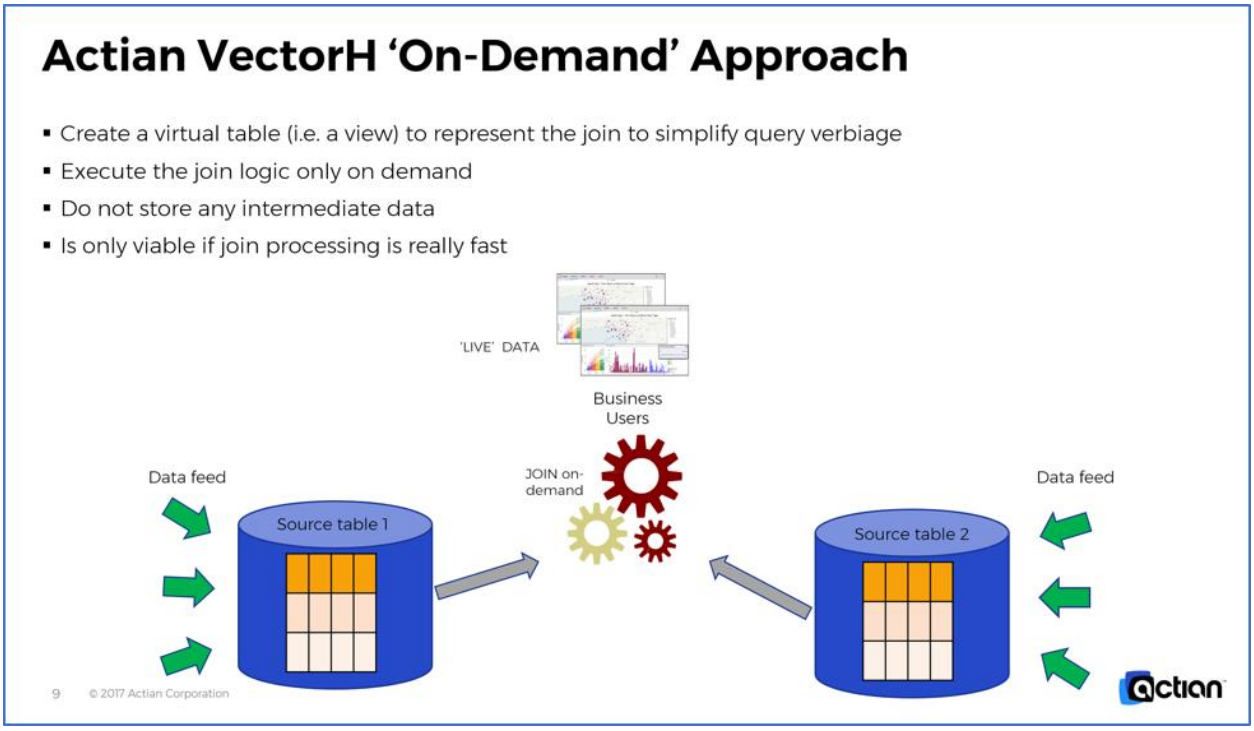
ACTIAN TOOLS SELECTION:

The users of the big data integrated warehouse have the need to combine the two very large tables in their databases, the fact and the dimension table. This is by far the best solution available for all the business questions from the users or customers. The legacy databases and Hadoop solutions have found difficult to combine the data as the data tables are large. They sometimes fail to finish.

The idea of materialization also failed as the query performance was brought down because of the huge data that the tables contained. Sub dividing the data was not an option as that leads to more concerns.

The solution provided the Actian tool was the Old School Materialization Test using the Actian VectorH, Actian's high performance columnar SQL database that runs in Hadoop was indeed an idea which would provide a faster and better solution. It was faster when compared with Hive. Not just the join but the entire process of materialization procedure in the same time frame as that of Hive.

The Actian On-Demand Approach demonstrated that with VectorH, queries against the base data without any intermediate materialization step, this was faster in a huge level that took every other method out of question.



In Conclusion the on-demand solution makes way for faster and more accurate results in real time.

References:

- https://searchdatamanagement.techtarget.com/feature/Integrating-data-warehouse-architecture-with-big-data-technology
- http://aircconline.com/ijcsit/V9N2/9217ijcsit01.pdf
- https://tdwi.org/Articles/2012/07/10/Big-Data-Staging-Area.aspx?Page=2
- https://www.researchgate.net/publication/298433319_Challenges_of_big_data_storage_and_management
- http://www.infotrellis.com/can-big-data-replace-edw/
- https://www.upwork.com/hiring/for-clients/big-data-testing-overcome-quality-challenges/
- https://mapr.com/blog/exploring-relationship-between-hadoop-and-data-warehouse-part-2/
- https://www.actian.com/company/blog/high-performance-realtime-analytics-hadoop-data/