

Big Data Implementation In Data Warehousing/Data Mart

Priyanka Kulkarni

Big Data : Overview - Tools and Use Cases

Alakh Verma

1 May 2016

Big Data Implementation In Data Warehousing/Data Mart

Introduction:

Codd paper published by E. F. Codd, of IBM's San Jose Research Laboratory (California 1970) has given a birth to concept of Relational Model of Data for Large Shared Data Banks. This white paper discusses about relational model, normal forms, redundancy, consistency and dependency of data. Relational database system is modelling of database consisting of collection of data in organized form for its retrieval, access and storage. This model is related to transactional data of any business. Relational database model which is intended and designed for analysing and querying of data is known as data warehouse. The name itself implies that it deals with huge amount of data along with analytics, data mining and decision making with that data. Its characteristics are subject oriented, integrated, non-volatile and time variant which makes this OLAP engine (ETL - Extraction, transportation, transformation and loading solution) different from the OLTP. Data Warehouse holds integrated data while Data Mart holds often one subject area. Data Mart is build in focus on dimensional model which can hold full but summarized data. In other way we can call it as subset of Data Warehouse and Big Data is nothing but solution to overcome limitations of Data Warehouse.

Relation of DBMS, Data Warehouse, Data Mart and Big Data:

Data Warehouse architecture with staging area and Data Mart can easily be visualize from below diagram.

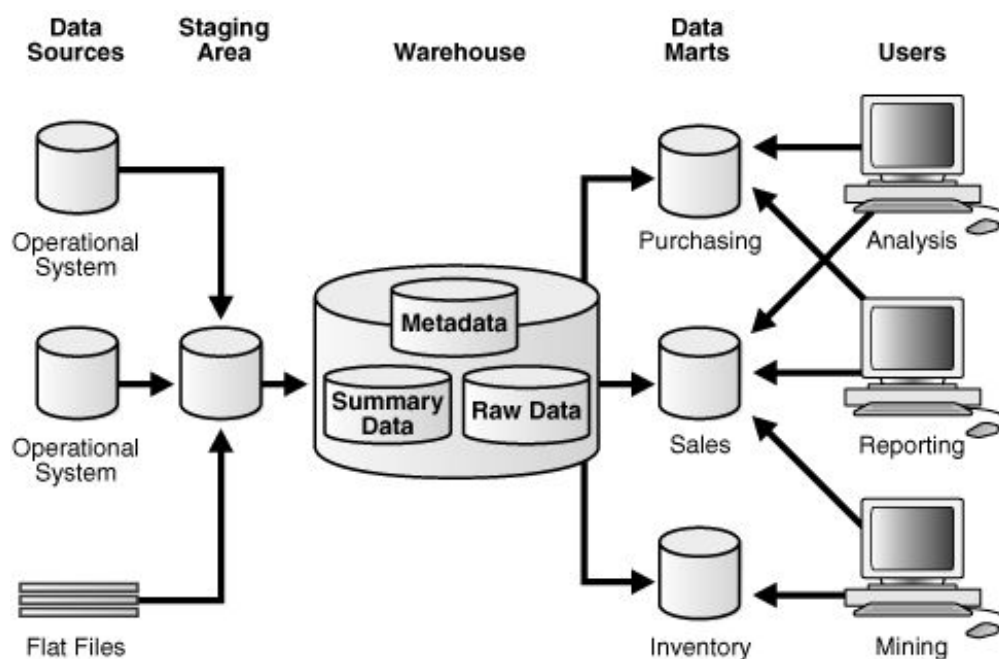


Figure 1. Architecture of Data Warehouse with staging area and Data Mart

Data Warehouse is architecture of data huge amount of data. Data Warehouse has its own strengths and limitations. To overcome limitations of unstructured data, it has to be handled predictably by analyzing it intelligently so actions can be taken to minimize expenses of storage devices. Big data is the solution for it. Hadoop is technology used by Big Data solutions which uses scale down approach and allows parallel processing by spanning data over different nodes using single cluster model. This help industries not only to analyze but also to arrange and use structured and unstructured data in a very skillful manner which turns to add value in the business areas. Datawarehouse and Datamart are used by many industries for their business intelligence reporting and dashboards.

Key Business Transformation:

Data Warehouse/Data Mart Infrastructure has undergone through a business transformation for advancement in its data modeling. Figure below shows that Big data can also be used as a data source for Data Warehouse which is more useful for industries using Data Warehouses since very long time. This data helps businesses to get more insight from warehouse applications.

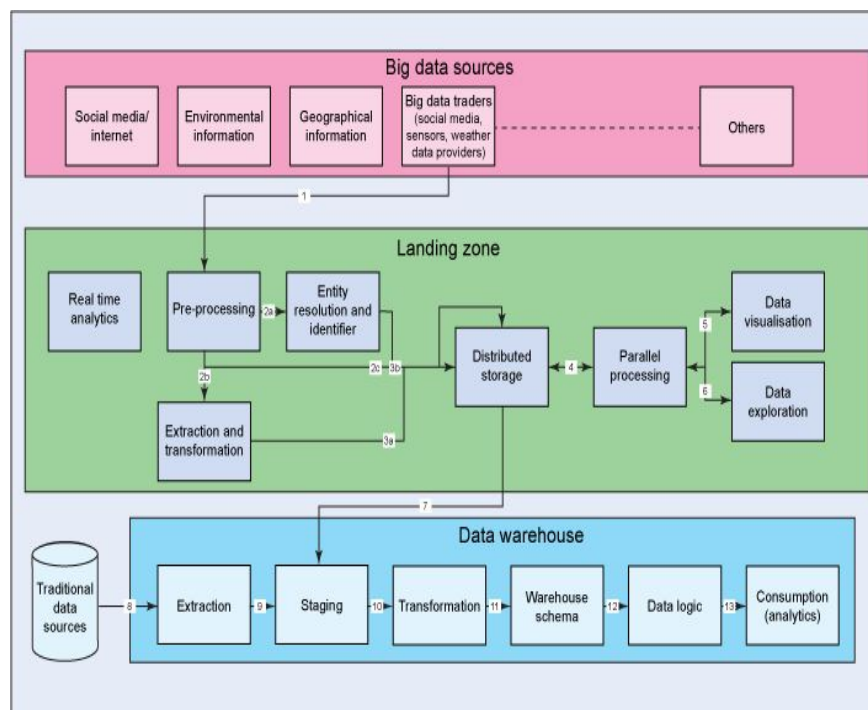


Figure 2. Big Data as a data source to the Data Warehouse

Big Data solutions have risen to fill an analytical gap in existing data models. Humanizing of existing volumouse data is essential and important for utilizing it effectively by Data artisan. Real world data keeps changing its volume as well as requirement. There are more than 15 different data warehousing methodologies available and its sources revolves

around three broad categories like infrastructure, core competency and information modeling. Some of its attributes are modeling of data, requirements, normalization of data, architecture designing, strategy implementation, managing metadata, query optimization and designing, data scalability and change in management.

Data Warehouse existing infrastructure can be modernized using big data technologies by extending its capabilities. Preprocessing, exploration and offloading are key types of this modernization. Using these methods organization can provide better query performance, optimize storage, deliver improved business insight to operations of organization for real time decision making, data warehouse rationalization with simplicity in low cost and combination of data.

Big Data Approach and Methodology:

Big Data methodologies and approaches provides effective solutions over Data Warehouses. Big data does processing on data and deploys its own solutions for improving it. It is very important to decide whether one should choose top down (Inmons Architecture) or bottom up (Kimball's Architecture) for transactions and analytics. Figure below shows tabular data of analytic competencies for managing data/information, analytic skills and tools to understand data and data oriented culture to act upon the data.

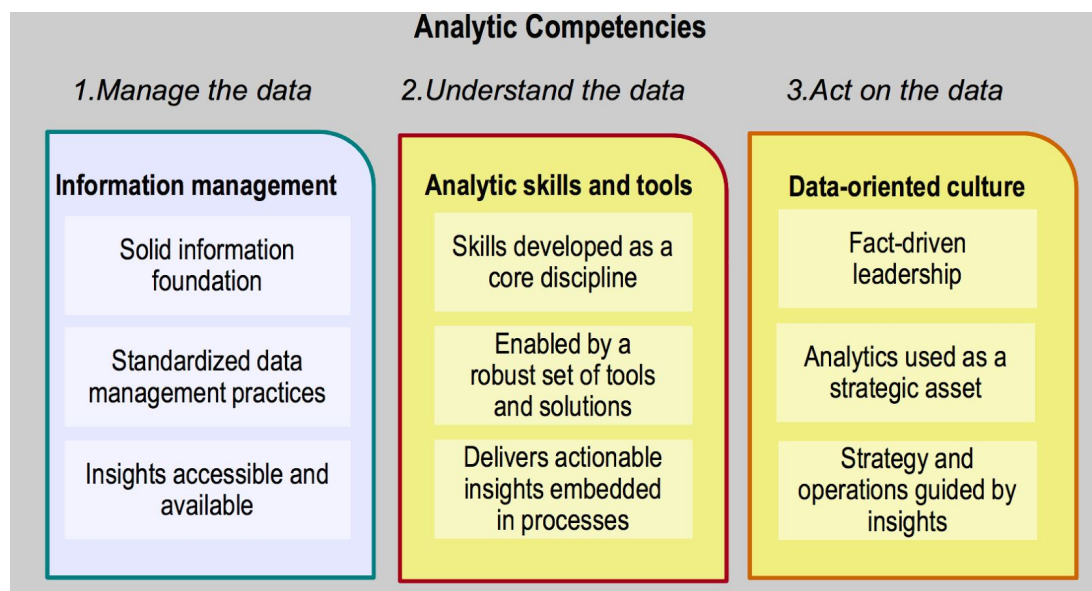


Figure 3. Analytic Competencies

Real world data in data warehouses is data with volume, variety and velocity. Current trend shows that managing data using big data solution is rapidly increasing. Data which is captured during many stages of operation is not very useful. There is a huge gap between use

of data and production of data. Figure below shows process of Hadoop (Big Data Framework) to interact with data in an effective manner.

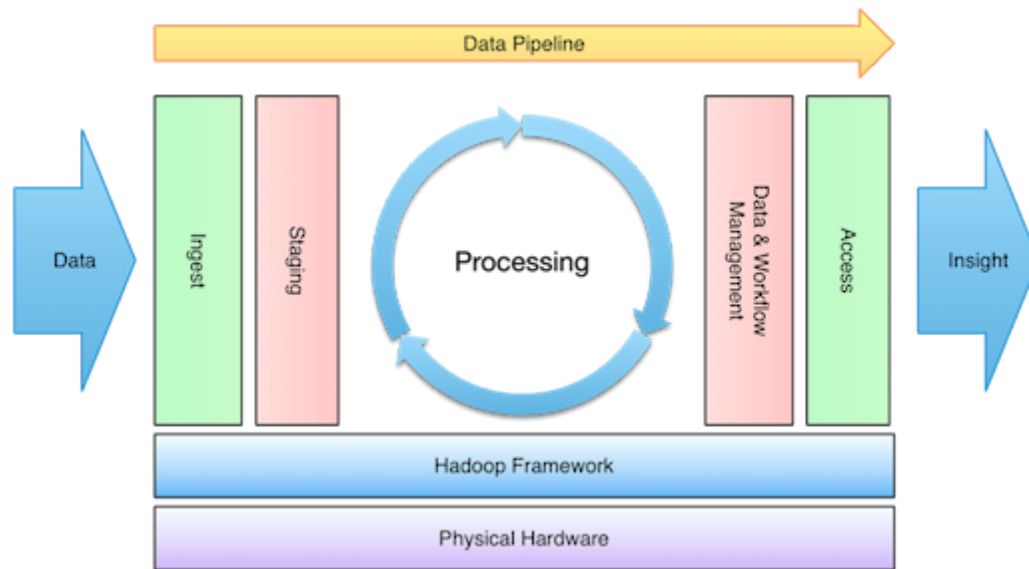


Figure 4. Extraction Of Valuable Data

An example of Hadoop is an implementation of mapReduce algorithm for HDLC which breaks down data into smaller fragments for parallel processing in distributed clusters. Data nodes communicate with each other for faster and efficient data generation and this architecture helps maintain data integrity. Each company can use their own Hadoop architecture, but data needs to be analysed to make required architectural changes.

Conclusion:

As technology advances, it is false to assume that previous technology or methods are deprecated. In other words new technology is born from old technologies, architectures and learning from old methodologies. It is not possible for all industries to adopt big data solutions. Usage of Datawarehouse and Datamart is important and unavoidable even if we come up with Big Data Solutions, all of them are connected like a tree if we consider database system as a seed, tree as data warehouse and fertilizers to grow that tree efficiently as big data solutions. These are essential to building a successful business enterprise. It is true that significance of data warehouse is reduced due to business intelligence and efficient big data solutions. Big Data has formalized path which helped data warehouse to move from its current level to the next one, but we need to keep monitoring sources for reaching to decisions from occurred events.

References

Cloudera CEO Mike Olson on Hadoop's architecture and its data applications

By James Turner January 12, 2011

No, Hadoop Isn't Going To Replace Your Data Warehouse

By Timo Elliott April 7, 2014

Data Warehouse vs Data Mart

By James Standen on 12/02/08

Big data Changing the way businesses compete and operate Insights on governance, risk and compliance April 2014

7 Steps to Data Warehousing

By David Walls and Mark D. Scott December 20, 1999

A Relational Model of Data for Large Shared Data Banks

By E. F. CODD IBM Research Laboratory, San Jose, California 1970

Oracle9i Data Warehousing Guide Release 2 (9.2) Part Number A96520-01

Data warehouse augmentation, Use big data technologies as a landing zone for source data

By Shweta Jain and Sujay Nandi 10 June 2014

How Big Data Technologies Can Augment and Complement Your DW

February 2013

Hadoop and the Data Warehouse: When to Use Which

By Dr. Amr Awadallah

Big Data Implementation vs. Data Warehousing

By Bill Inmon November 7, 2013

Towards Business Intelligence over Unified Structured and Unstructured Data Using XML

By Zhen Hua Liu and Vishu Krishnamurthy

Course Material of Big Data: Overview- Tools and Use cases (UCSC Extension)

By Alakh Verma