Big data survey and application

Seminar on Big Data Management
Tsegaye Beka Ture
Presentation Outline

- Introduction
- Big data management
- Big Data technologies
- Data analytics
- Adding Big Data capability to an existing information system
- Identifying relevant pieces of information in messy data
- Summary and conclusion

Seminar on Big Data Management   Tsegaye Beka Ture
Introduction

Why BigData? Why not just data?

- The data is too big, moves too fast, or doesn’t fit the structures of your database architectures and infrastructures.
- Big data is data that is too big to be handled and analyzed by traditional database protocols such as SQL (Structured Query Language).
Introduction

Why BigData? Why not just data?

- Big Data may evolve over time, what is now big data may quite rapidly may become small. In this sense, size is just one aspect of these new technologies.
- This fast growth of data is due to advances in digital sensors, communications, computation, and storage that have created huge collections of data.
Introduction

Why BigData? Why not just data?

- The volume, variety and velocity of available information exponentially increase the profound complexity and enormity of information that companies need to manage and these factors generate questions they haven’t previously encountered in the course of doing business.

- The data generated by people and the system in the internet nowadays is measured in exabytes ($10^{18}$) and zettabytes ($10^{21}$).
Big data includes structured data, E-mail, Word file, picture, chat record, video, music and other information.
Introduction

Why BigData ? Why not just data?

- The world population is more than 7 billion in the mean time and out of this more 2 billion are connected to the internet and more than 5 billion people use mobile devices.
- By 2025, the forecast is that the internet will exceed the brain capacity of everyone living on the entire planet.
- Dealing effectively with Big Data requires one to create value against the volume, variety and veracity of data while it is still in motion (velocity), not just after it is at rest.
Big Data Management

Big Data Survey involves in data generation, acquisition, storage and analysis or processing:

- Generation and acquisition
- Storage and organization
- Processing and analysis

**Acquisition**

Big Data architecture has to acquire high speed data from a variety of sources (web, DBMS(OLTP), NoSQL, HDFS) and has to deal with diverse access protocols. It is where a filter could be established to store only data which could be helpful or “raw” data with a lower degree of uncertainty.
Global Information Storage Capacity
in optimally compressed bytes

1986
ANALOG
2.6 exabytes

1993

2000

Digital Storage

2002:
“beginning of the digital age”

50%

2007
ANALOG
19 exabytes

- Paper, film, audiotape and vinyl: 6%
- Analog videotapes (VHS, etc): 94%
- Portable media, flash drives: 2%
- Portable hard disks: 2.4%
- CDs and minidisks: 0.8%

DIGITAL

- Computer servers and mainframes: 8.9%
- Digital tape: 11.8%
- DVD/Blu-ray: 22.8%
- PC hard disks: 44.5%
- 123 billion gigabytes
- Others: < 1% (incl. chip cards, memory cards, floppy disks, mobile phones, PDAs, cameras/camcorders, videogames)


% digital:
1 %
3 %
25 %
94 %

Digital
280 exabytes
Big Data Management

Organization/Storage

At this point the architecture has to deal with various data formats (texts formats, compressed files, variously delimited, etc.) and must be able to parse them and extract the actual information like named entities, relation between them, etc. Also this is the point where data have to be clean, put in a computable mode, structured or semi-structured, integrated and stored in the right location (existing data warehouse, data marts, Operational Data Store, Complex Event Processing engine, NoSQL database).
Big Data Management

Analyze/Process
Running queries, modeling, and building algorithms to find new insights. Data mining requires integrated, cleaned, trustworthy data; at the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions.
Big Data Management

Decision

Being able to take valuable decisions means to be able to efficiently interpret results from analysis. Consequently it is very important for the user to “understand and verify” outputs.

Furthermore, provenance of the data (supplementary information that explains how each result was derived) should be provided to help the user to understand what he obtains.

If we can easily see how volume, velocity, veracity and variety influence the pipeline of Big Data architecture, there is another important aspect in data to handle: privacy.
Big Data Management

The various aspects of Big Data are clustered into four main domains

1) **Technical Data Provisioning** which includes acquisition, storage and processing

2) **Technical Data Utilization** which includes computation and time complexity

3) **Functional Data Provisioning** which includes information life cycle, lean information management, value oriented information management etc

4) **Functional Data Utilization** which includes the realms where big data is used
Proposed Big Data lifecycle

- Retrieve/reuse/discover
  - Searching
  - Retrieval
  - Decision making

- Security
  - Privacy
  - Confidentiality
  - Integrity
  - Availability
  - Governance

- Sharing/publishing
  - Ethical and legal specification
  - Organization and documentation
  - Representation

- Storing
  - Management plans
  - Content filtering
  - Distributed system
  - Partition tolerance
  - Consistency

- Big Data
  - Raw data

- Collection
  - Cleaning/integration
    - Log files
    - Sensing
    - Mobile equipment
    - Satellite
    - Laboratory
    - Supercomputer

- Filtering/classification
  - Structure/unstructure
  - Cleaning/integration
  - Filtering criteria

- Data analysis
  - Visualization/interpretation
  - Technique and technology
  - Tool selection
    - Data mining algorithm
    - Cluster
    - Correlation
    - Statistical
    - Regression
    - Legacy codes
    - Indexing
    - Graphics
Big Data Management

Handling Big Data implies having an infrastructure linear scalable, able to handle high throughput multi-formatted data, fault tolerant, auto recoverable, with a high degree of parallelism and a distributed data processing. It is important to note that, in this management, integrating data (i.e. “access, parse, normalize, standardize, integrate, cleanse, extract, match, classify, mask, and deliver data.”) represents 80% of a Big Data project.
Big Data technologies

There are various tools which can be used in Big Data management from data acquisition to data analysis.

Most of these tools are parts of Apache projects and are constructed around the famous Hadoop.

Written in Java and created by Doug Cutting, Hadoop brings the ability to cheaply process large amounts of data, regardless of its structure.
Hadoop is made up of two core projects: **Hadoop Distributed File System (HDFS)** and **MapReduce**. HDFS is a distributed file system designed to run on large clusters of commodity hardware based on Google File System (GFS).
HDFS and MapReduce involved Big Data Technologies are classified according to their capabilities

- **Storage and Management Capability**
  - **Cloudera Manager**: an end-to-end management application for Cloudera’s Distribution of Apache Hadoop.
  - **RCFile (Record Columnar File)**, a data placement structure for structured data. Here, tables are vertically and horizontally partitioned, lazily compressed. It is an efficient storage structure which allows fast data loading and query processing.
HDFS and MapReduce Big Data technologies

- **Database Capability**
  - **Oracle NoSQL** a high performance <key, value> pair database convenient for non-predictive and dynamic data thus for Big Data;
  - **Apache HBase** a distributed, column-oriented database management system, modeled on Google’s Big Table that runs on top of HDFS
  - **Apache Cassandra** a database which combines the convenience of column-indexes and the performance of log-structured updates
  - **Apache Hive** can be seen as a distributed data warehouse. It enables easy data ETL from HDFS or other data storage like HBase or other traditional DBMS. It has the advantage of using a SQL-like syntax, the Hive QL;
  - **Apache ZooKeeper** is an open-source, in-memory, distributed NoSQL database that is used for coordination and naming services for managing distributed applications
HDFS and MapReduce Big Data technologies

- **Processing Capability**
  - **Pig** which is intended to allow people using Hadoop to focus more on analyzing large datasets and thus spend less time having to write mapper and reducer programs
  - **Chukwa** which is a data collection system for monitoring large distributed systems
  - **Oozie** which is an open-source tool for handling complex pipelines of data processing. Using Oozie, users can define actions and dependencies between them and it will schedule them without any intervention
HDFS and MapReduce Big Data technologies

- Data Integration Capability
  - **Apache Sqoop**: a tool designed for transferring data from a relational database directly into HDFS or into Hive. It automatically generates classes needed to import data into HDFS after analyzing the schema’s tables; then the reading of tables’ contents is a parallel MapReduce job.
  - **Flume**: is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data. It is designed to import streaming data flows.
Big data Technologies

The 10 hottest big data technologies:

- **Predictive analytics**: software and/or hardware solutions that allow firms to discover, evaluate, optimize, and deploy predictive models by analyzing big data sources to improve business performance or mitigate risk.

- **NoSQL databases**: key-value, document, and graph databases

- **Search and knowledge discovery**: tools and technologies to support self-service extraction of information and new insights from large repositories of unstructured and structured data that resides in multiple sources such as file systems, databases, streams, APIs, and other platforms and applications.
Big data Technologies

- **Stream analytics**: software that can filter, aggregate, enrich, and analyze a high throughput of data from multiple disparate live data sources and in any data format.

- **In-memory data fabric**: provides low-latency access and processing of large quantities of data by distributing data across the dynamic random access memory (DRAM), Flash, or SSD (Solid State Drive) of a distributed computer system.

- **Distributed file stores**: a computer network where data is stored on more than one node, often in a replicated fashion, for redundancy and performance.
**Big data Technologies**

- **Data virtualization**: a technology that delivers information from various data sources, including big data sources such as Hadoop and distributed data stores in real-time and near-real time.

- **Data integration**: tools for data orchestration across solutions such as Amazon Elastic MapReduce (EMR), Apache Hive, Apache Pig, Apache Spark, MapReduce, Couchbase, Hadoop, and MongoDB.

- **Data preparation**: software that eases the burden of sourcing, shaping, cleansing, and sharing diverse and messy data sets to accelerate data’s usefulness for analytics.

- **Data quality**: products that conduct data cleansing and enrichment on large, high-velocity data sets, using parallel operations on distributed data stores and databases.
Big data Technologies

What is Big Data?

Mammoth the NullSQL tool
New Breakthrough in Big Data Technologies: the NullSQL Paradigm shift
Mammoth the NullSQL tool

- The NullFS, or the **Null File System** contains a **name node** which keeps track of a set of **Data nodes** in a **cluster formation**.

- Each of these **data nodes** contain a **data storage device** that is **write-only**, the write only device has the advantage of being **to hold any amounts of data you write onto it**. This device is the equivalent of achieving **quantum computing in big Compute**. It is the underlying breakthrough for the **NullSQL** paradigm.
Big Data Technologies
Mammoth the NullSQL tool

- The Architecture of Mammoth is similar to Hadoop, it is portable, distributed, fault tolerant, and scalable.
- More importantly, it does satisfy all 3 aspects of the CAP theorem, Concurrency, Availability, and Partition tolerance.
- It has two layers, the NullFS Layer, and the MapRemove layer.
- The MapRemove layer will optimally copy data to the right device that is free.
- The device drivers for both Windows and Unix: /dev/null (UNIX) and NUL (Windows) are implemented fully by Windows Azure.
- The Job tracker and task tracker can optimally handle batch processing jobs as well as jobs submitted by different users.
Big Data Technologies
Mammoth the NULLSQL tool
Cloud Clients
Web browser, mobile app, thin client, terminal emulator, ...

SaaS
CRM, Email, virtual desktop, communication, games, ...

PaaS
Execution runtime, database, web server, development tools, ...

IaaS
Virtual machines, servers, storage, load balancers, network, ...
The Big Data Open Source Technology Stack

- Search and Query languages
  - SQL
  - Hive

- Databases
  - Cassandra
  - CouchDB
  - MySQL
  - MongoDB

- Data Warehouses
  - Hadoop

- File Systems
  - HBase

- Storage hardware/Data Centers

Created by Pravi Solutions
Big Data analytics

Big Data Analytics can be defined as the use of advanced analytic techniques on big data. Nowadays, we can put big data and analytics together. There are many techniques used for analytics on big data

- **Association rule learning** to find relationships among entities (mainly used in recommendation systems).
- **Machine learning** to bring computer to learn complex patterns and make intelligent decisions based on it
- **Data mining** which can be seen as a combination of statistics and machine learning and statistics with database management
Big Data analytics

- **Cluster analysis** used as unsupervised machine learning. It aims to divide data into smaller clusters having the same set of characteristics not known in advance.

- **Crowdsourcing** used to collect data and/or features and metadata to enhance the current semantics of data.

- **Text analytics** which aims to analyze large text collections (email, web pages, etc.) to extract information. It is used for topics modeling, question answering, etc.
Big Data Analytics Infrastructure: Technical or IT Buy

**Data Stack**
- Structured & Unstructured
- In-Memory Appliances
- Database Appliances
- Enterprise Data warehouses
- Local Data warehouses
- Datamarts ODS

**Hadoop and Big Data Ecosystem**
- Search and Visualization (Lucene)
- Data Orchestration (Hbase, Flume, ZooKeeper)
- Data Access (Pig, Hive, Sqoop, Avro)
- Hadoop (HDFS, MapReduce)
- NoSQL Databases – MongoDB, Couch

**Enterprise Information Management Stack**
- Data Governance
- Data Integration
- Data Quality
- Data Virtualization
- Master Data Management

**BI Platforms, Analytics Tools and Insight Stack**
- Machine Learning (pattern discovery)
- Predictive (Forecasting, Recommendations)
- Prescriptive (Simulation, What-if)
- Descriptive (Statistics, Historical)
- Reporting Scorecards Dashboards

**BI Platforms, Data and Analytics is the #1 investment area for corporations. But the amount of innovation, themes and paths are causing confusion among managers**
Adding Big Data capability to an existing information system

- Data integration capability needs to deal with velocity and frequency.
- The challenge here is also about ever growing volume and, because many technologies leverage Hadoop, use technologies that allow you to interact with Hadoop in a bidirectional manner: load and store data (HDFS) and process and reuse the output (MapReduce) for further processing.
Adding Big Data capability to an existing information system

• The main challenge is not to build a system that is ideally suited for all processing tasks but to have an underlying architecture flexible enough to permit to processes built on top to work at their full potential.

• There is not a commonly agreed solution, an infrastructure is intimately tied to the purpose of the organization
Identifying relevant pieces of information in messy data

As mentioned in the semantics challenges list, this task can be done before the disambiguation pile. In this case, we must prune irrelevant data. This pruning is mostly done by a “bag of words” approach. It helps through cosine similarity to rapidly compare things (documents and sentences) and to select relevant one, according to a threshold.

If done after the pile, this task can be seen as the build of a Big Data index. Obviously, this problem is mainly broached by people who intend to design a search engine.
Summary and Conclusion

- Big Data is a term coined in order to explain data that is too big, moves too fast, has different forms, and uncertainty which enables the user to derive important value to make critical decisions based on the result.
- Big Data is evolving and dynamically growing, therefore volume is just one aspect of the different properties of Big Data.
- Big Data analytics involves different phases from generation to acquisition, from storage and organization to analysis and processing and then making decision.
- The important aspect of Big Data is the refined end result that helps companies to make important decision for better performance of their business.