Big Data

Presented By Group E
Stepan Mincev, Artsiom Hramyka, Chun Paul Ho , Siqi Chen and Yani Hu

Agenda Layout

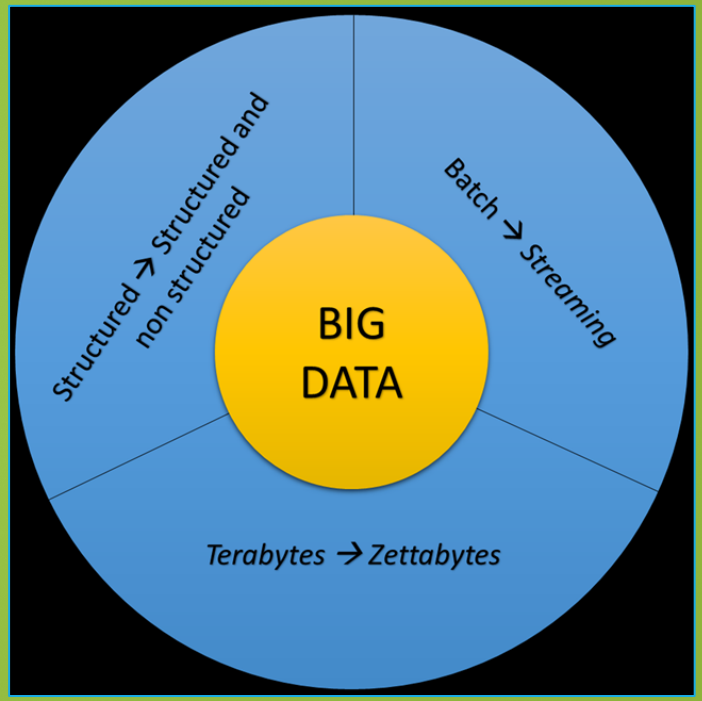
- Big Data Value Chain
- Tools
- Application
- Limitations
- Further Research

History of Big data, what is Big Data?

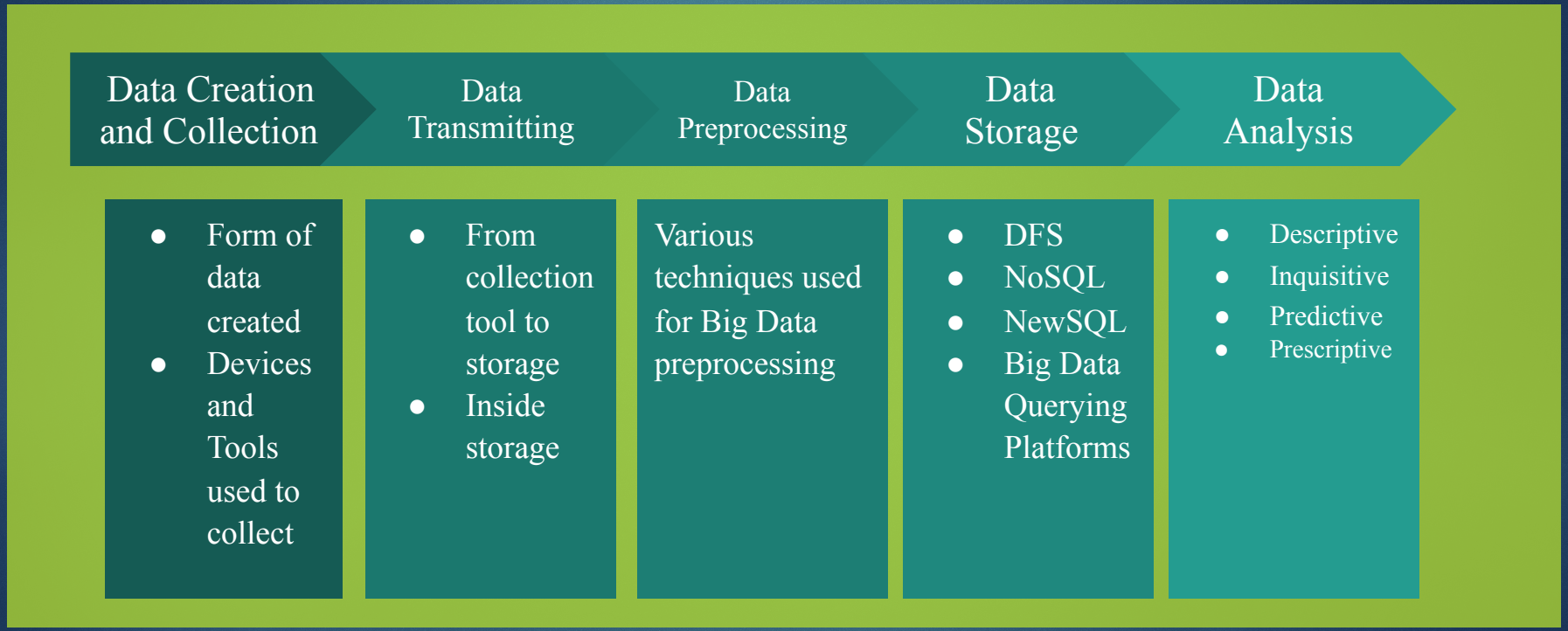
- Summary of **literature review**, states that the definition of Big Data refers to datasets which are **too grand** for typical software tools to capture, store, manage and analyse.
- The original notion of Big Data dates back to **2001**, where the challenges of increasing data addressed with a 3V's model which was later developed into the 6Vs model.

Dimensions of Big Data- from 3Vs to 6Vs

- Volume
  - Magnitude of data
- Velocity
  - Rate of generation of data
- Variety
  - Different types of generated data
- Veracity
  - Unreliability associated with data sources
- Variability
  - Variation in the flow rate of data
- Low-Value density
  - Unusable data



Big Data Value Chain



Tools for collecting, pre processing and analysing big data

- Advancements in technologies help managers in scenario building analysis
- Relational Database Management System (RDBMS) is the traditional method
- Hadoop Distributed File System (HDFS)
  - Fault-tolerant, scalable, highly configurable distributed storage system

Other file systems include Apache Cassandra, Apache Hive and others

Software tools for handling big data

Many new languages, frameworks and data storage technologies have emerged which support the handling of big data:

- R
  - Open source statistical computing language, provides a wide variety of statistical and graphical techniques to derive insights from data
- Python
  - Open source and supported by Windows, Linux and Mac platforms
- Scala
  - Increasingly growing programming tool for handling big data problems
- Few other frameworks that support Big data include:
  - Apache Spark, Apache Hive, Apache Pig, Amazon Elastic Compute Cloud (EC2), MongoDB, BlinkDB, Tachyon, Cassandra, CouchDB and others

NoSQL databases

Redis

An open-source in-memory data structure project implementing a distributed, in-memory key-value database with optional durability.

Yahoo!
Facebook

MongoDB

A Document store NoSQL database. The data is stored in the form of a document . JSON

the City of Chicago, Codecademy, Foursquare, IBM, The Gap, Inc., Uber.

Cassandra

Cassandra is a hybrid, non-relational database similar to Google's BigTable.

The main clients are twitter and facebook.

Other NoSQL databases include CouchDB and others

Application of Big Data Analytics

The concept of Big Data analytics has left no sector untouched. Here are merely a few various sectors and how they use big data analytics:

- Healthcare
  - Multiple sources are used to gain insights, these include electronic patient records; clinical decision support system; clinical data; and machine generated sensor data.
- Telecommunication
  - To improve customer experience, Mobile service providers analyse a number of factors such as demographic data, customer preferences, household structure and usage details.
  - Network Analytics is the next big thing in Telecom, where MSP's can monitor network speed and manage the entire network
- Financial firms
  - Capital markets are using big data in preparation for regulations like EMIR, Solvency II, Basel II etc.
  - The timeliness of finding value plays an important role in both investment banking and capital markets, hence, there is a need for real-time processing of data.

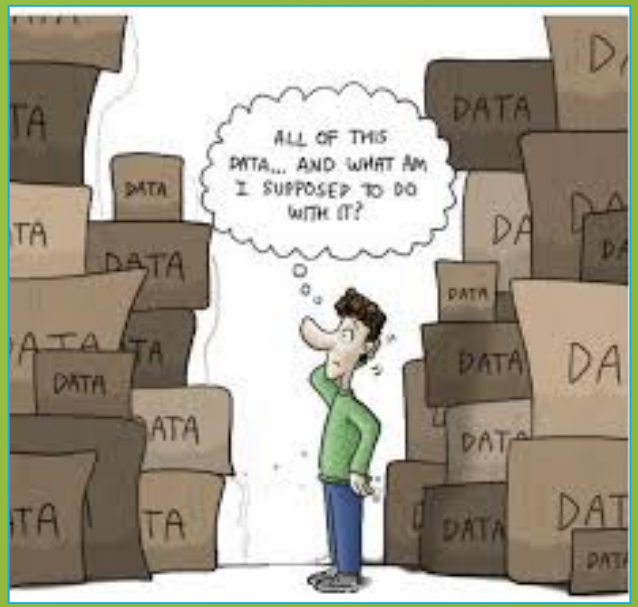
Key focus on Retail (E-Commerce)

- Retail
  - Market Basket Analysis- buyers of certain products are more inclined to purchase other specific items
  - End goal is to achieve customer value

Activity data	Clickstream Data	Video Data	Voice Data
Amazon engages a type of predictive modeling technique called collaborative filtering, using customer data to generate 'you might also want' prompts for each product bought or visited.	eBay Inc. conducts thousands of experiments with different aspects of its website to determine optimal layout and other features ranging from navigation to the size of its photos	Some retailers utilize sophisticated image-analysis software linked to their video-surveillance cameras to track in-store traffic patterns and consumer behavior	Credit card companies use and track call center activities to make personalized offers in milliseconds and to optimize offers by tracking responses

Problems with Big Data

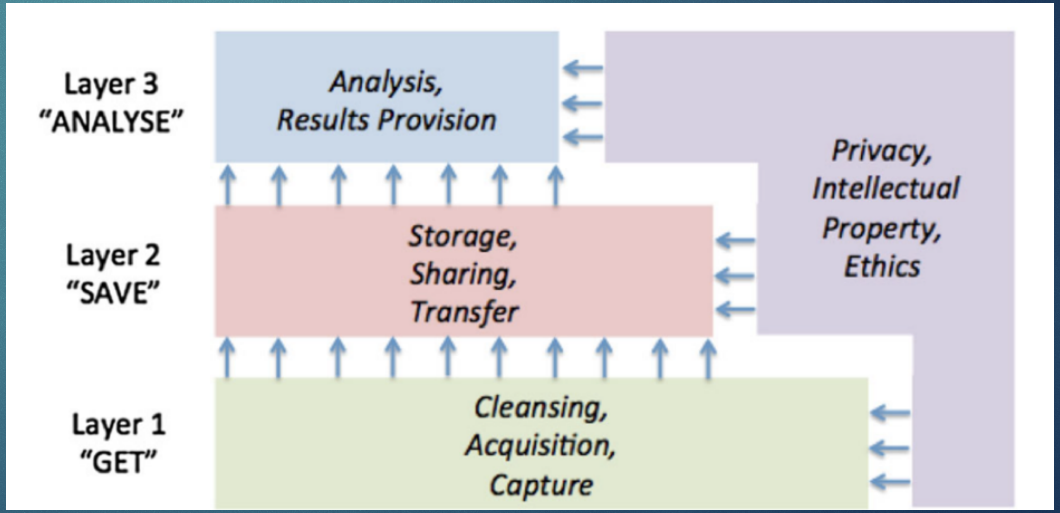
- Data Privacy
- Data Quality (Data Diversity, Data volume, Data Change & Lack of consensus)
- Data Security
- Data Discrimination
- Data Ownership
- Data Security
- Scalability
- Data Storage and Transportation



Further Research

Future research challenges could be discussed according to the basic fundamental data process :

Four layers and five aspects(Anagnostopoulos et al., 2016).



Further Research

1. Effective data cleaning and life cycle of data
2. Storage and dynamic data placement
3. Develop algorithms and Scalability software platforms
4. Ethical considerations

Conclusion

- Big data becoming more and more important for firms across a plethora of sectors
- Especially with at least 60% boost in operating margin
- Big Data handling will be become routine for many
- Large firms may develop their own Big Data processing facilities and hire expert specialists
- Small firms may have to rely on large companies to subcontract their Big Data requirements
- All companies however need to consider the challenges mentioned as there is a recurring trend that Companies are unable to stay ahead of the dynamic nature of Big Data in terms of processing capabilities