Challenges and Opportunities with Big Data

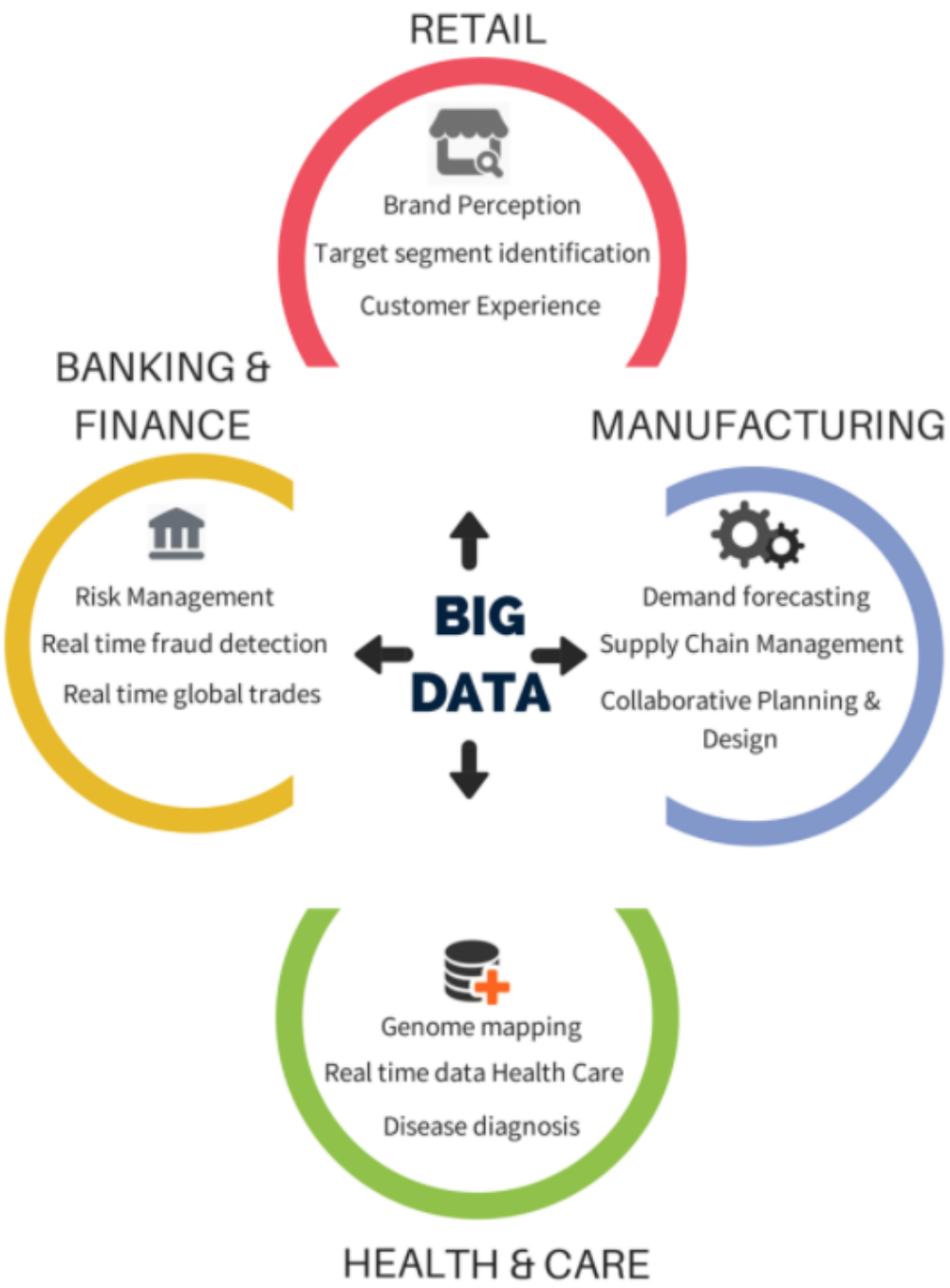
By: Rohit Ranjan

Introduction

- **What is Big Data?**
- **Big data** is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them (Wiki Definition)
- Usually data >1Terabytes is considered Big-Data
- In 2010, 13 Exabyte (13*1024*1024 Terabytes) of new data was generated by users and enterprises
- Potential value of only global personal location is more than $700 Billion

Why is Big Data Important ?

- It can help us with:
  - Cost reductions for industry and users
  - Time reductions,
  - New product development and optimized offerings, and
  - Smart decision making
  - New Research in Medical, Life Sciences, Astronomy
  - Fraud Detection, Security


RETAIL
Brand Perception
Target segment identification
Customer Experience
BANKING & FINANCE
Risk Management
Real time fraud detection
Real time global trades
BIG DATA
MANUFACTURING
Demand forecasting
Supply Chain Management
Collaborative Planning & Design
Genome mapping
Real time data Health Care
Disease diagnosis
HEALTH & CARE

Why do we have Big-Data Problem?

- Too Many bytes (Volume)
- Too high a rate (Velocity)
- Too many sources (Variety)

- Non Scalable Source
  - Human-Intensive
  - Compute Intensive

Big Data Analysis Pipeline

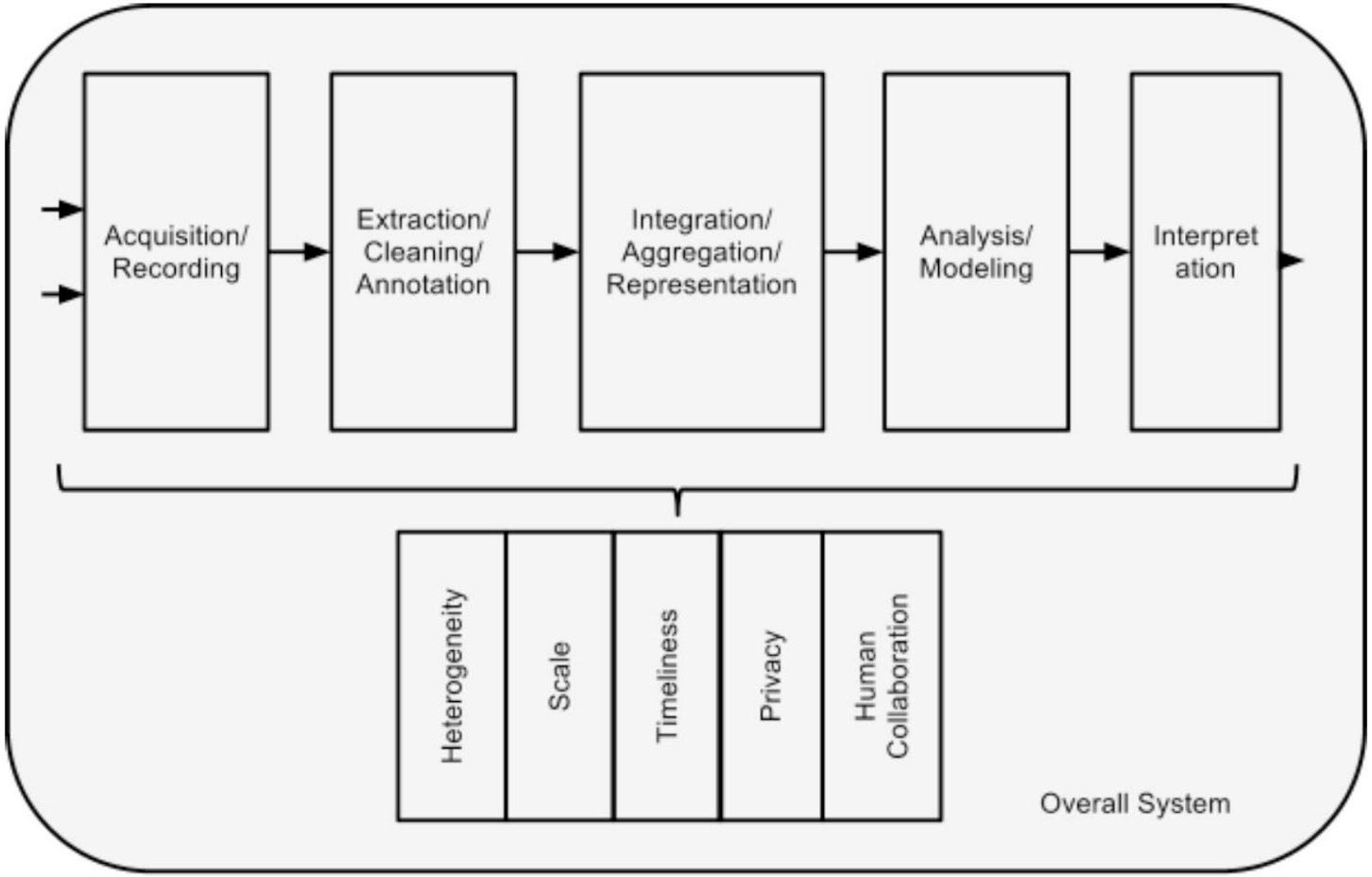


Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

Phases in Processing Pipeline

- Data Acquisition and Recording
- Information Extraction and Cleaning
- Data Integration, Aggregation, and Representation
- Query Processing, Data Modeling, and Analysis
- Interpretation

Data Acquisition and Recording

- Mainly Two challenges in Data Acquisition:
  - Define filters in such a a way that discard non-useful data and do not discard any useful data
  - Automatically generate the right metadata to describe what data is recorded, how it is recorded and measured

Information Extraction and Cleaning

- Information collected will not be in a format ready for analysis
- We require an information extraction process that pulls out required information and express it in structured format

Data Integration, Aggregation, and Representation

- Storing, Identifying, and forming right meta-data should happen in an automated manner (i.e intelligent database design)
- We must enable professionals to create effective effective database design, so that even in absence of intelligent database design, they can create effective databases

Query Processing, Data Modeling, and Analysis

- Querying and Mining Big Data are different from traditional statistical analysis
- It's noisy, dynamic, heterogeneous and untrustworthy
- Problem with Big –Data analysis is lack of coordination between database systems
- Tight coupling between database system will benefit expressiveness and analysis

Interpretation

- Ability to analyze Big-Data is of limited value, if users can't understand
- Rich palette of visualization is important in conveying users, the results of queries in a way, that's best understood in particular domain

Challenges in Big Data Analysis

- Heterogeneity and Incompleteness
- Scale
- Timeliness
- Privacy
- Human Collaboration

Heterogeneity and Incompleteness

- Machine expect homogenous data, and cannot understand nuance
- Even after data cleaning, and error correction some incompleteness and error in data remains

Scale

- Data Volume is scaling faster than compute resources
- Computing are shifting to parallel processor, so now we have to deal with parallelism within single node
- Cloud Computing: It aggregates multiple disparate workloads, with varying performance goals
- Storage devices: Shift from HDD to Solid State drives, it requires rethinking of how we design storage subsystem

Timeliness

- Design a system that can process given size of data faster
- Example: In credit card fraud we should be able to detect fraud in timely manner.

Privacy

- Privacy is huge concern in today's big-data system
- Design a system that can give user fine-grained control over sharing

Human Collaboration

- Many patterns are easily detected by humans but computer algorithms have hard time finding it
- We can incorporate human and machine together to get a better visualization of data

System Architecture

- New kind of System Architecture is needed to process Big-Data
- Big-Data Typically involves multiple phases (from gathering data, extraction, data cleaning, statistical modelling etc.)
- Current system offer little to no support for Big-Data Pipeline

Thank You!!

Question and Answer?