XORIANT

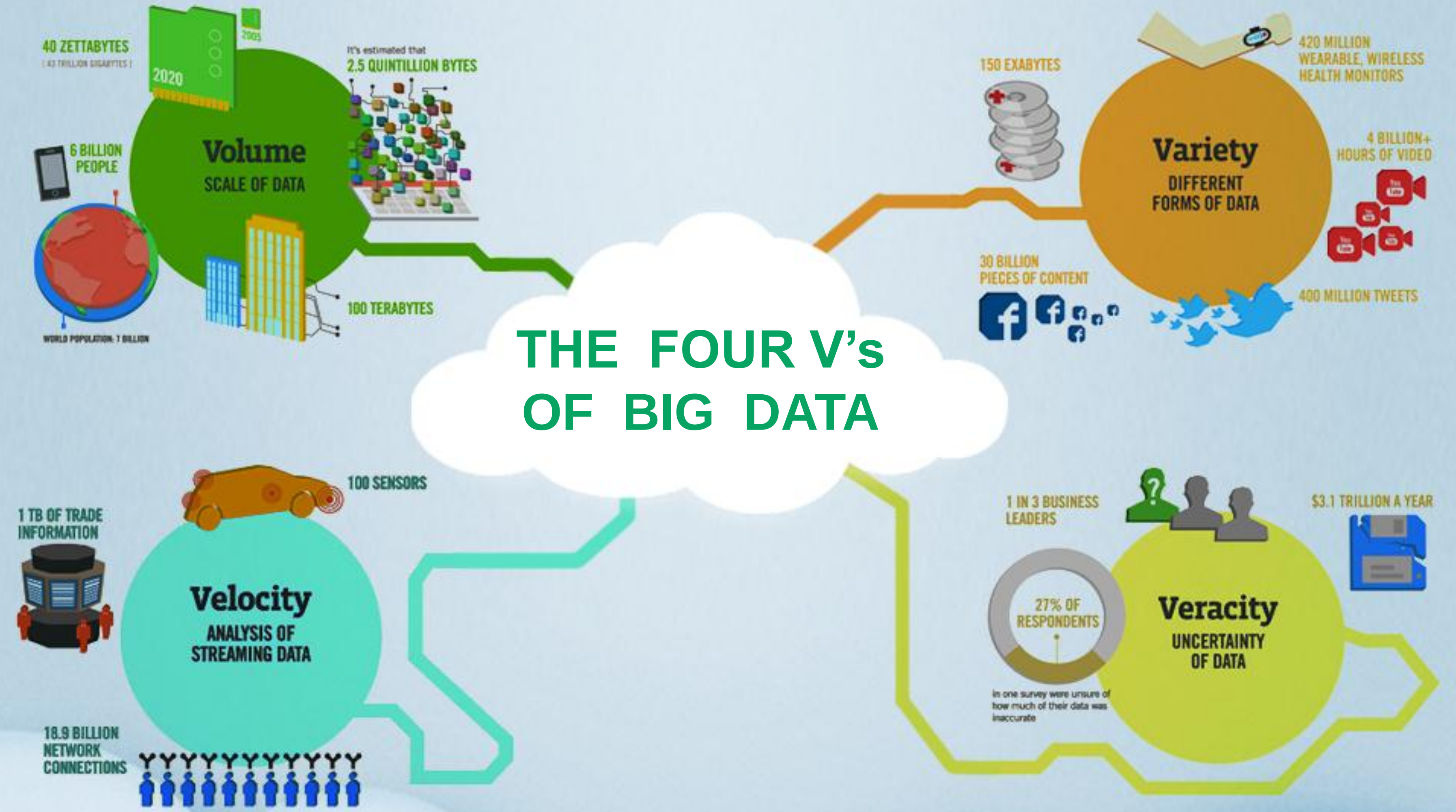
# Addressing Big Data Challenges: The Hadoop Way

Presented by: Atul Dambalkar

# Agenda

- Big Data Challenges
- Big Data Analytics Industry Trends
- Hadoop as a Solution
- Real Life Solution Studies
  - Case Study I - Retail Industry
  - Case Study II - Online Advertising Industry
- How Xoriant can help?
- Q & A

# Big Data Challenges


THE FOUR V's
OF BIG DATA
40 ZETTABYTES
( 43 TRILLION GIGABYTES )
2020
2005
It's estimated that
2.5 QUINTILLION BYTES
6 BILLION PEOPLE
Volume
SCALE OF DATA
100 TERABYTES
WORLD POPULATION: 7 BILLION
150 EXABYTES
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
Variety
DIFFERENT FORMS OF DATA
4 BILLION+ HOURS OF VIDEO
30 BILLION PIECES OF CONTENT
400 MILLION TWEETS
1 TB OF TRADE INFORMATION
100 SENSORS
Velocity
ANALYSIS OF STREAMING DATA
18.9 BILLION NETWORK CONNECTIONS
1 IN 3 BUSINESS LEADERS
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA

*Source: IBM*

# Traditional Approach & Its Limitations

## Data Warehousing Vendors (ETL)

INFORMATICA
The Data Integration Company
TERADATA
SAP
talend*
open data solutions
sas
NETEZZA
ORACLE
EXADATA
pentaho
open source business intelligence

Costs - High Initial Setup, Maintenance, Subscription or Licensing Fees

No support for ad-hoc query

No support for unstructured data

Multiple copies of data in different formats

Data latency and bottlenecks

*Note: The Logos are proprietary of the individual companies*

# Big Data and Analytics - Trends

## ETL Trends

Enterprise Data Hub or Data Lake (Hadoop w/ HDFS)

- Open Source Software
- Commodity Hardware
- No multiple data copies
- Fault-Tolerant storage for Raw data As-Is
- Current Limitations - Write/Append only, No Delete or Update

## Data Processing Trends

- Unified Data Access
- Multiple Data Processing Paradigms
- In-memory processing
- In-memory, Real-time Stream processing
- Analytics based on Distributed SQL Processing

## Architecture Trends

- Falling memory prices
- Batch mode processing for Data Size  more than Hundreds of TBs
- In-memory processing for Data Size less than Hundreds of TBs

# Hadoop Proposition

Open Source Ecosystem

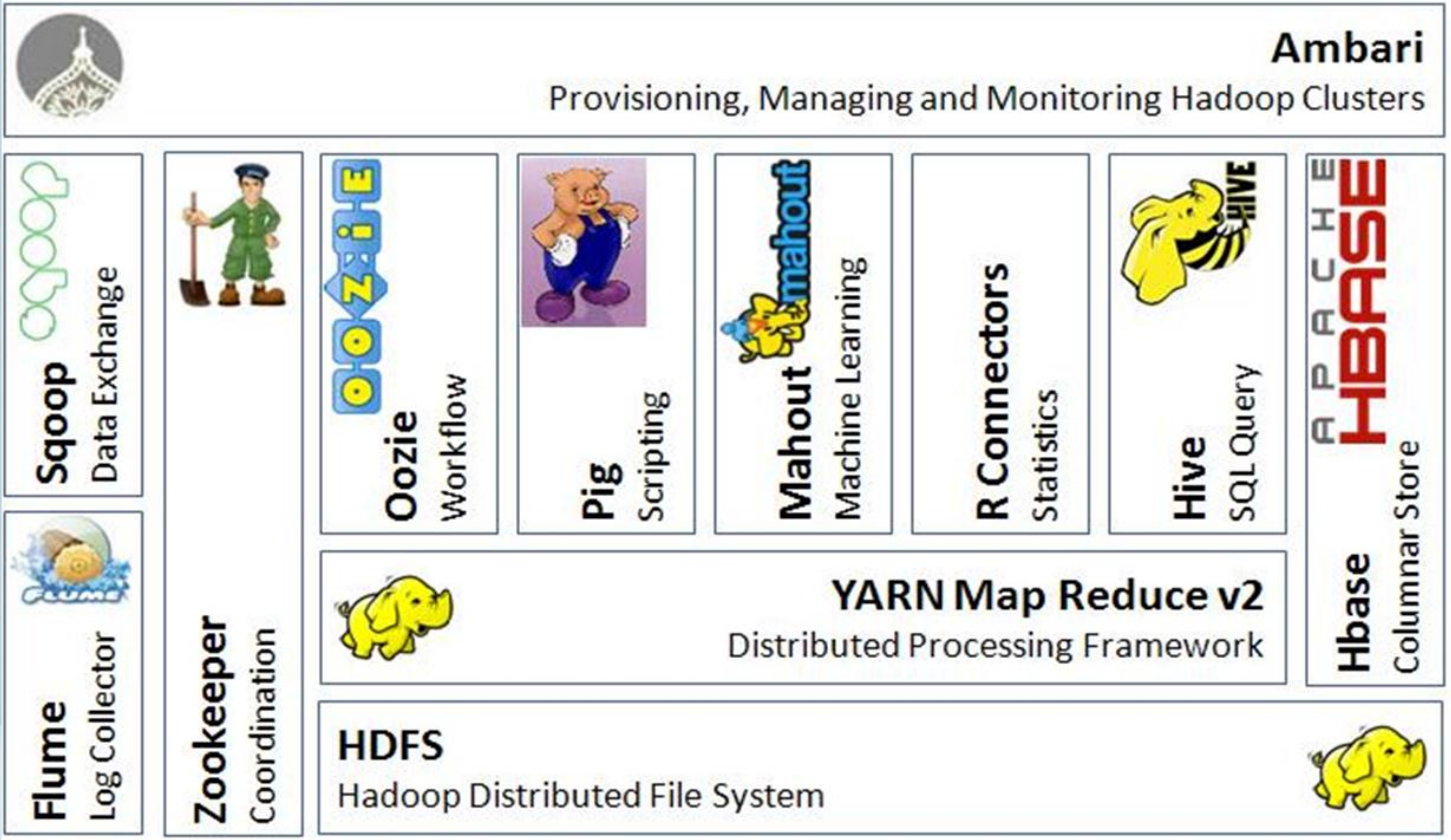
No Data Loss through replicated storage (HDFS)

Runs on commodity hardware

Multiple data analysis/processing paradigms

- Map-Reduce
- Script based (Pig Latin)
- SQL like - HiveQL, Apache Drill, Presto (Facebook),
- Impala (Cloudera), HAWQ (Pivotal)
- In-memory Processing (Apache Spark)

# Apache - Hadoop Ecosystem



**Ambari**
Provisioning, Managing and Monitoring Hadoop Clusters

**Sqoop**
Data Exchange

**Flume**
Log Collector

**Zookeeper**
Coordination

**Oozie**
Workflow

**Pig**
Scripting

**Mahout**
Machine Learning

**R Connectors**
Statistics

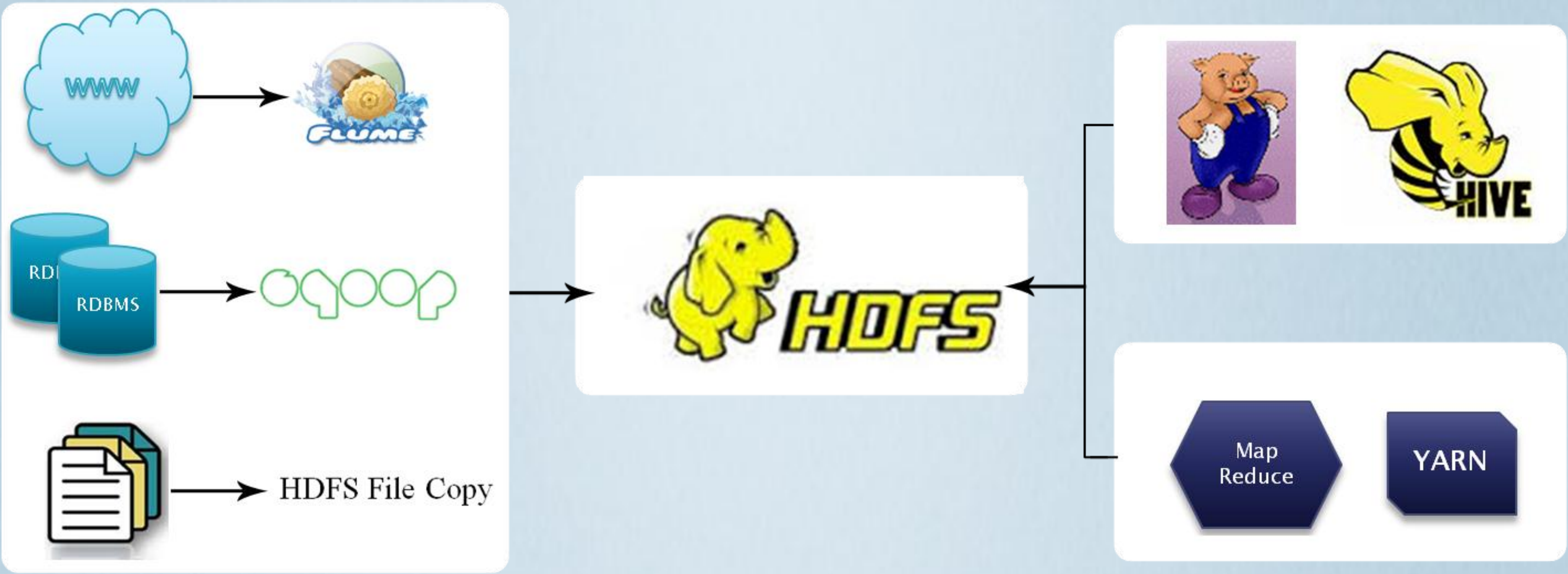
**Hive**
SQL Query

**Hbase**
Columnar Store

**YARN Map Reduce v2**
Distributed Processing Framework

**HDFS**
Hadoop Distributed File System

# Hadoop Data Flow



www

FLUME

RDBMS

RDBMS

sqoop

HDFS File Copy

HDFS

HIVE

Map Reduce

YARN

Data Loading

Data Storage

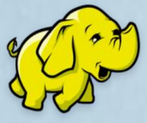
Data Analysis

# Case Study - 1

## Retail Industry

### Problem Scenario

- Personalize marketing campaigns, coupons, offers, marking down inventories
- Improving customer loyalty – leads to sales and profitability
- Competition from other retailers
- ETL based analysis tasks - taking lot of time – up to 6 weeks
- Software systems (Oracle, Greenplum, SAS, Teradata)
- Mainframe based expensive hardware systems



### Hadoop based Solution

- Data stored into HDFS with replication
- 300 Hadoop nodes with 2PB data

### Benefits

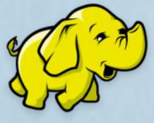
- Data processing time down to 1 week and even daily
- Mainframe cost savings
- No software licensing Costs
- Limitless data storage with HDFS
- No multiple data copies
- Low cost

# Case Study - 2

## Online Advertising Industry (Attribution Computation)

### Problem Scenario

- Growing Ads Impression and conversion events
- Longer attribution computation time (6 to 8 hours for each computation run). Advertisers needed quick results
- Unable to process more than 150GB data within each computation
- IBM Netezza based solution along with Oracle
- Expensive hardware and software costs



### Hadoop based Solution

- Data Stored into HDFS with replication
- Initially used HiveQL then moved to Cloudera Impala (MPP architecture based  Distributed SQL Engine)

### Benefits

- Attribution computation time down to 45 minutes
- Capable of processing up to 300GB data for each computation
- Manageable data storage with HDFS
- Low cost

# Xoriant Big Data Practice - Overview

- Understands technological needs and organizational challenges faced with respect to Big Data
- Understands rapidly evolving Big Data technology space
- Can help bridge the gaps with Big Data capabilities
- Brings Big Data and NoSQL technology expertise

# Thank you!

## Do you have any Questions?

**Xoriant – Big Data Center of Excellence**

Email: bigdata@xoriant.com

**For FREE consultation, please contact us on the above mentioned email address.**