# Census Data Analysis using Big Data Technology

PREPARED BY:- ANNU SHARMA

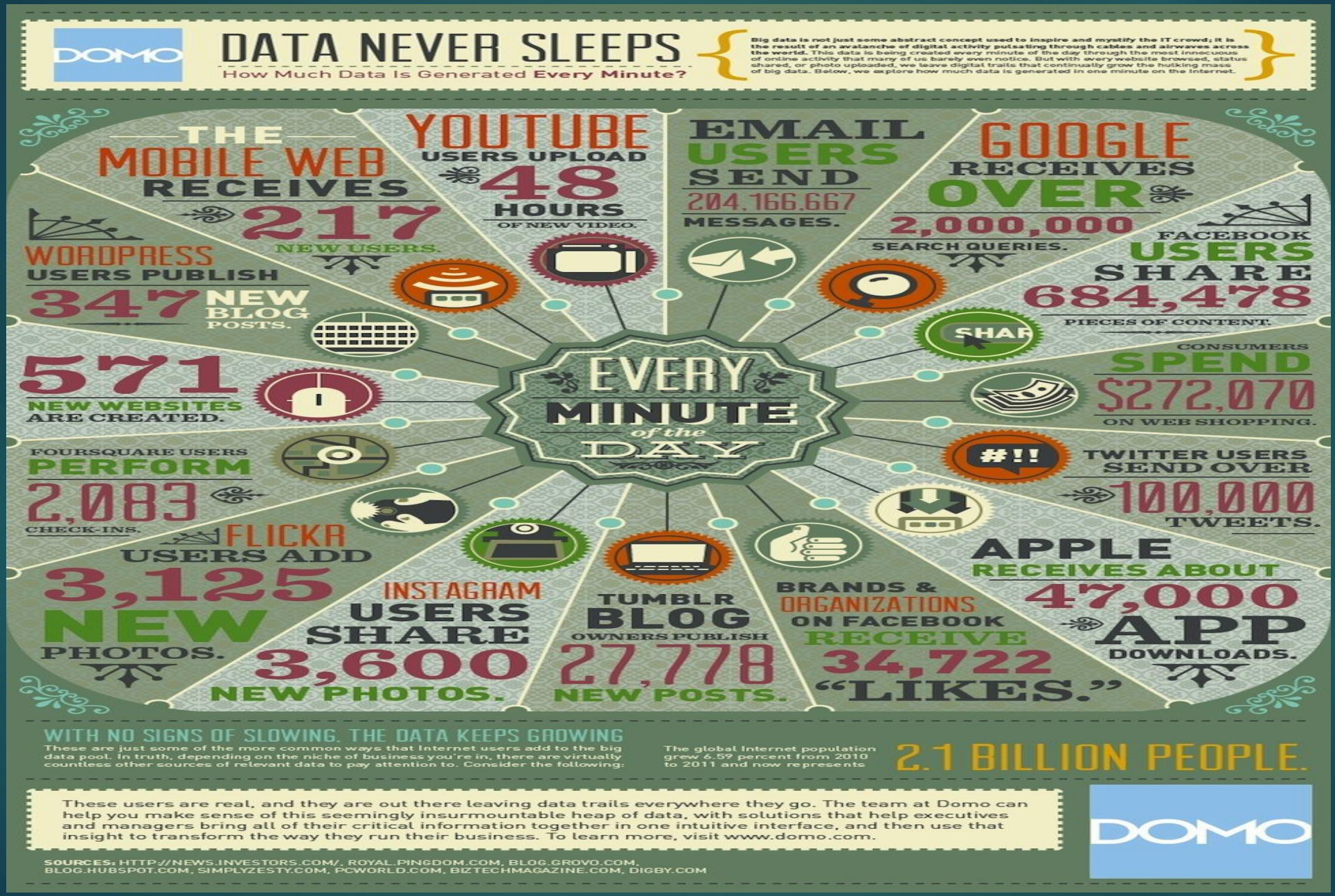
# Content

- The Context : Big Data
- What is Big Data
- Big Data Use Cases
- Characteristics Of Big Data
- Hadoop Eco-System
- Importance of Big Data Solutions
- Project Outline
- Architectural Diagram
- Hardware and Software Requirements
- Job1
- Job2
- Job3
- Job4
- Questions
- Thankyou

# The Context : Big Data

- Man on the moon with 32KB (1969); my laptop had 2GB RAM (2009)
- Google collects 270PB data in a month (2007), 20000PB a day (2008)
- New York Stock exchange generates huge amount of data in a day

- Data mining huge amounts of data collected in a wide range of domains from astronomy to healthcare has become essential for planning and performance.
- We are in a knowledge economy.
  - Data is an important asset to any organization (Social, Retail, Transport, Technology)
  - Discovery of knowledge; Enabling discovery; annotation of data
- Business Success Heavily depends upon
  - Ability to store and process huge amount of data
  - We are looking at newer
    - programming models, and
    - Supporting algorithms and data structures.


DOMO
DATA NEVER SLEEPS
How Much Data Is Generated Every Minute?
Big data is not just some abstract concept used to inspire and mystify the IT crowd; it is the result of an avalanche of digital activity pulsating through cables and airwaves across the world. This data is being created every minute of the day through the most innocuous of online activity that many of us barely even notice. But with every website browsed, status shared, or photo uploaded, we leave digital traits that continually grow the hulking mass of big data. Below, we explore how much data is generated in one minute on the Internet.
THE MOBILE WEB RECEIVES 217 NEW USERS.
YOUTUBE USERS UPLOAD 48 HOURS OF NEW VIDEO.
EMAIL USERS SEND 204,166,667 MESSAGES.
GOOGLE RECEIVES OVER 2,000,000 SEARCH QUERIES.
FACEBOOK USERS SHARE 684,478 PIECES OF CONTENT.
WORDPRESS USERS PUBLISH 347 NEW BLOG POSTS.
571 NEW WEBSITES ARE CREATED.
EVERY MINUTE of the DAY
CONSUMERS SPEND $272,070 ON WEB SHOPPING.
FOURSQUARE USERS PERFORM 2,083 CHECK-INS.
TWITTER USERS SEND OVER 100,000 TWEETS.
FLICKA USERS ADD 3,125 NEW PHOTOS.
INSTAGRAM USERS SHARE 3,600 NEW PHOTOS.
TUMBLR BLOG OWNERS PUBLISH 27,778 NEW POSTS.
BRANDS & ORGANIZATIONS ON FACEBOOK RECEIVE 34,722 "LIKES."
APPLE RECEIVES ABOUT 47,000 APP DOWNLOADS.
WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING
These are just some of the more common ways that Internet users add to the big data pool. In truth, depending on the niche of business you're in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:
The global Internet population grew 6.59 percent from 2010 to 2011 and now represents 2.1 BILLION PEOPLE.
These users are real, and they are out there leaving data trails everywhere they go. The team at Domo can help you make sense of this seemingly insurmountable heap of data, with solutions that help executives and managers bring all of their critical information together in one intuitive interface, and then use that insight to transform the way they run their business. To learn more, visit www.domo.com.
DOMO
SOURCES: HTTP://NEWS.INVESTORS.COM/, ROYAL.PINGDOM.COM, BLOG.GROVO.COM, BLOG.HUBSPOT.COM, SIMPLYZESTY.COM, PCWORLD.COM, BIZTECHMAGAZINE.COM, DIGBY.COM

# Big Data Use-Case

### Sentiment Analysis

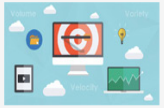
Sentiment analysis offers powerful business intelligence to enhance the customer experience, revitalize a brand, and gain competitive advantage. The key to successful sentiment analysis lies in the ability to mine multi-structured data pulled from a variety of sources into a single database. Learn how a big data platform can help you get more value out of sentiment analysis.

### 360-Degree Customer View

A 360-degree customer view offers a deeper understanding of customer behavior and motivations. Obtaining a 360-degree customer review requires analysis of data from sources like social media, data collecting sensors, and mobile devices. From there, more effective micro-segmentation and real-time marketing often result. Learn why big data analytics offers a more complete picture of the consumer.

### Ad-hoc Analysis

Ad-hoc analysis only looks at the data requested or needed, providing another layer of analysis for data sets that are becoming larger and more varied. Big data ad-hoc analytics can help in the effort to gain greater insight into customers by analyzing the relevant data from unstructured sources, both external and internal. Learn how about a how big data cloud service makes ad-hoc analysis easier in Hadoop.



### Real-time Analytics

Systems that offer real-time analytics quickly decipher and analyze data sets, providing results even as data is being generated and collected. This high-velocity method of analytics can lead to instant reaction and changes, allowing for better sentiment analysis, split testing, and improved targeted marketing. Learn more about how your business can benefit from real-time analytics.

### Multi-Channel Marketing

Multi-channel marketing creates a seamless experiences across different types of media like company websites, social media, and physical stores. Successful multi-channel marketing requires an integrated big data approach during all stages of the buying process. Learn more about how a big data platform can streamline your multi-channel marketing.


### Customer Micro-Segmentation

Customer micro-segmentation provides more tailored and targeted messaging for smaller groups. This personalized approach requires analysis of large sets of data collected through customers' online interactions, social media, and other sources. Learn more about how companies are using big data to better segment and target customers.

### Ad Fraud Detection

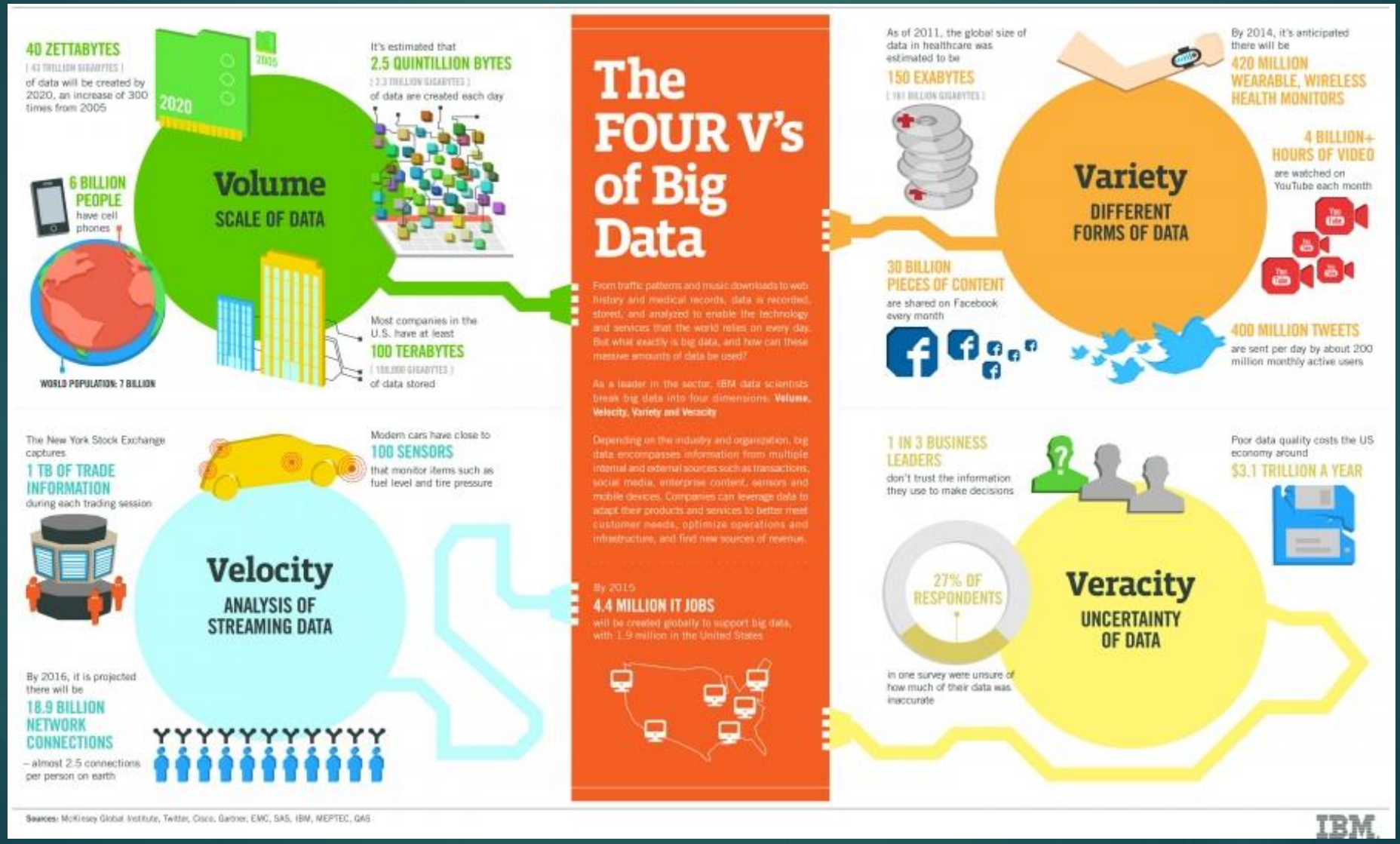
Ad fraud detection requires data analysis of current fraud strategies by recognizing patterns and behaviors. Data that shows abnormalities of group behavior make it so ad fraud is detected early and stopped before it is spread. Learn why businesses are turning to big data platforms to combat ad fraud.

### Clickstream Analysis

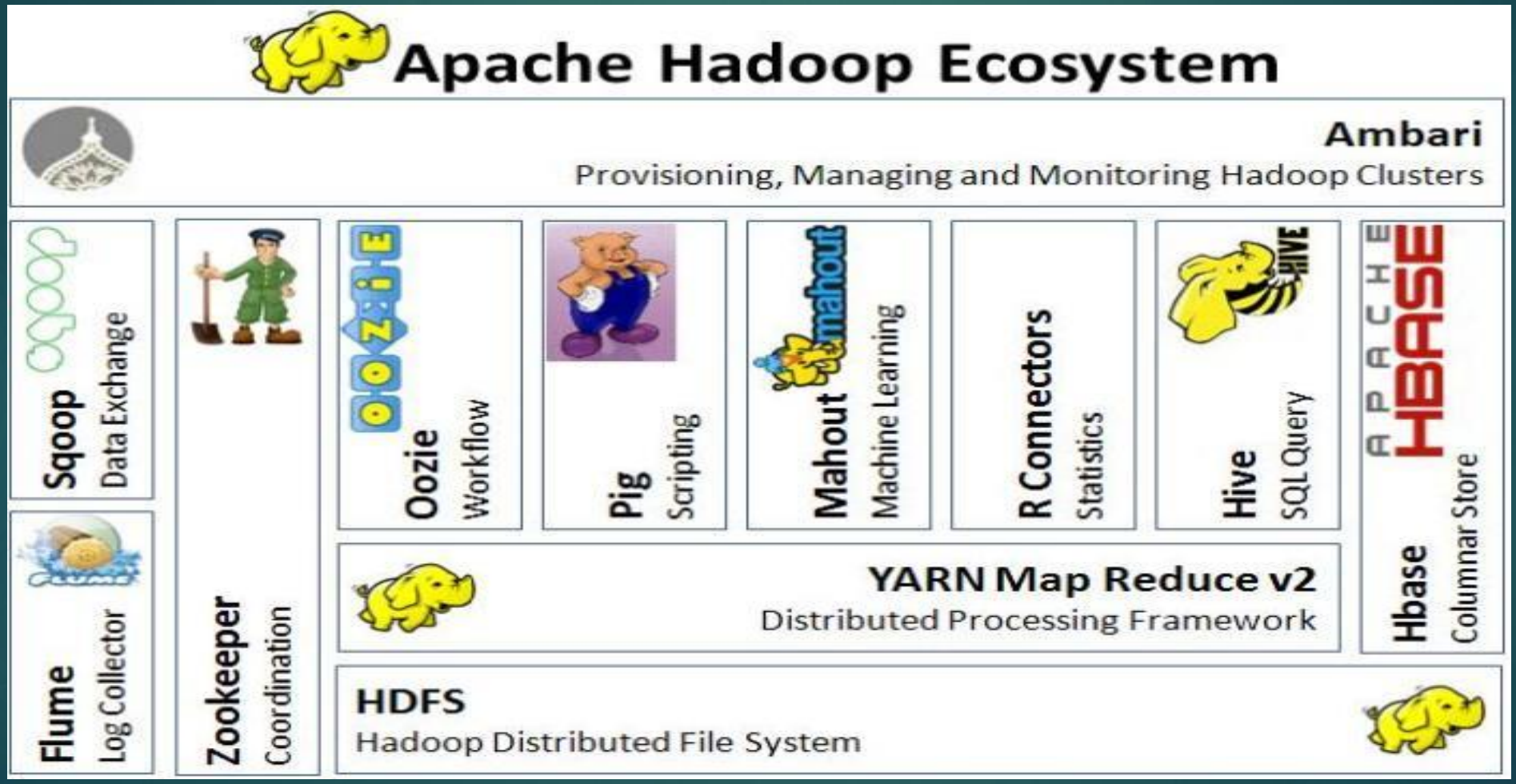
Clickstream analysis helps to improve the user experience by analyzing customer behavior, optimizing company websites, and offering better insight into customer segments. With big data, click stream analysis helps to personalize the buying experience, getting an improved return on customer visits. Learn more about the impact of big data on clickstream analysis.




# Characteristics Of Big Data



# Hadoop Eco-System



# Importance of Big Data Solutions

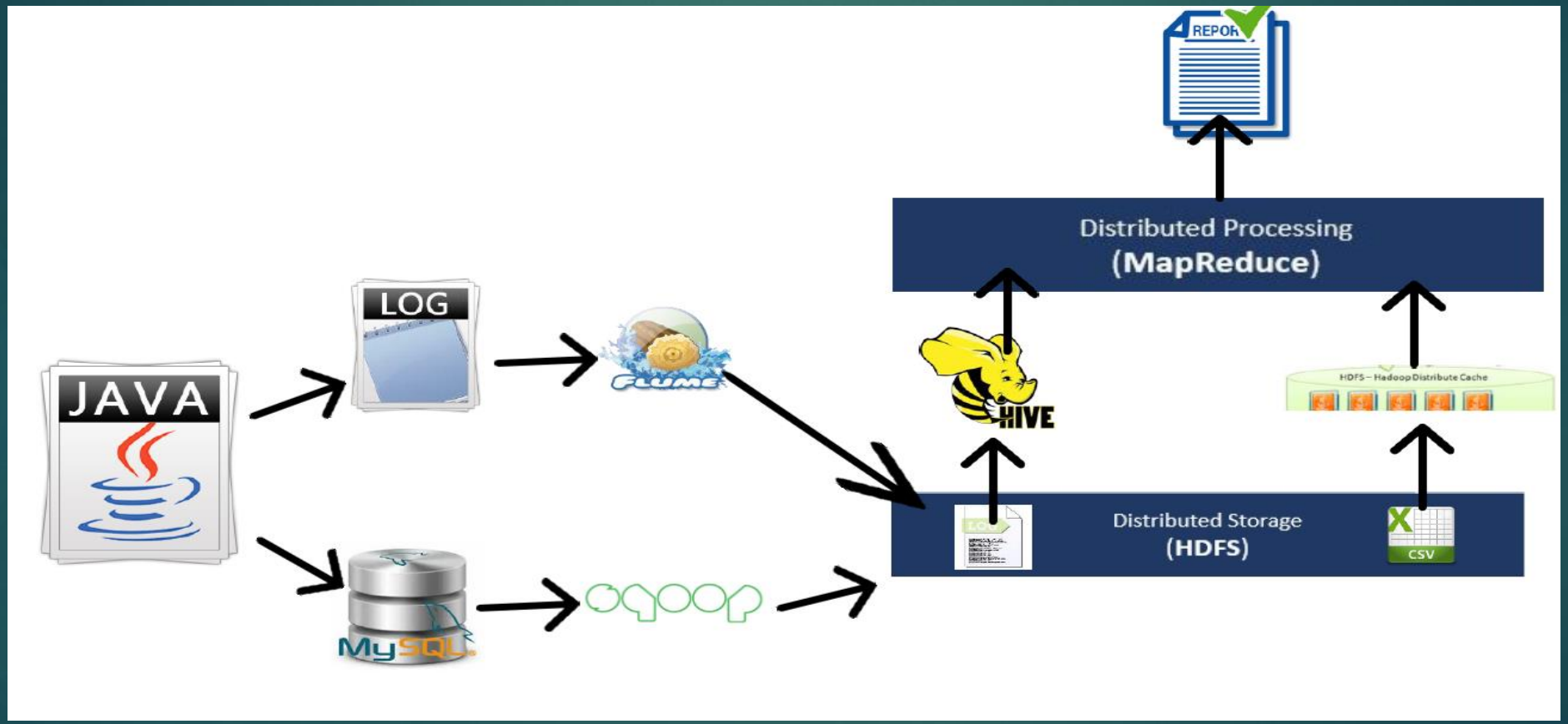
# Project Outline

| Title | Census Data Analysis Using BigData Hadoop |
|---|---|
| Inputs | Raw Census Data in JSON format & supporting Data Table |
| Data Elements | Age, Education, Marital Status, Gender, TaxFiler Status, Income, Parents, Country of Birth, Citizenship, WeeksWorked |
| Analysis Relevance | Educational , Social, Financial and Planning |
| Purpose | To help various organizations including government in reforming the life of citizens by launching various schemes related with Education, Social, Economical. It will also help the planning commission. |
| Methodology | SCRUM and Agile |

# Project Outline Cont..

- In this project we are aims at using the Hadoop ecosystem to Analyse a Census Data. The data is accepted from user based upon their choice. The raw data is provided in the form of Json logs. Supporting data is also maintained using MySQL database. The Hadoop Ecosystem is used to work upon the big data extracted from the Json log files and MySQL database together and extract substantial amount of information like the country's Literacy Rate, Orphan percentage, Average Income, Taxpayers details, Male & Female Ratio.
- The complete process includes data import from server logs to HDFS using Flume and cleaning up of this data using Hive. Also master data for age group mapping will be downloaded using Sqoop from MySQL. The final steps would be generating reports using core MapReduce jobs using Distributed Cache and Map-Side / Reduce-Side Join, PIG Latin design pattern.

# Architectural Diagram



# Hardware & software Requirement

- Hardware :-
  - 8 GB Ram
  - Quad Core Processor
  - 100 GB HD
- Software :-
  - Virtual Box
  - Ubuntu & Cloudera virtual .ova file
  - Hadoop , HDFS, PIG, Hive, Scoop, Flume, MySQL, Java, TomCat

# Job 1 : Educational Analysis

- Task 1 : Calculate :
  - Education Qualification Count : Sub-grouped by Gender
  - Education Qualification Count based on Employment
  - Calculate Sex Ratio (Male : Female)

- *No Supporting Table Required*

# Job 1 : Educational Analysis

- Final Output
  - Task1
    - ( Children, Male,228)
    - ( Children, Female,224)
    - ( 9th grade, Male,27)
  - Task2
    - ( 9th grade,25)
    - ( 10th grade,37)
    - ( 11th grade,42)
    - ( 5th or 6th grade,12)
  - Task3
    - ( Male,939)
    - ( Female,1061)

# Job 2 : Financial Analysis

- **Task 1 : Calculate :**
  - **Tax based Income Generated**
    - Total Income Generated , Gender wise Total Income Generated
    - Total Tax Payers
    - Total Tax to be collected
  - **Per Capita Income Analysis**
    - Per Capita Income
    - Age Group wise Per Capita Income
    - Gender wise Per Capita Income
- **Supporting Tables Required :**

| Column Name | Column Type | Remarks |
|---|---|---|
| Gender | Varchar | Store gender info |
| minincome | int | Stores min salary of a salary slab |
| maxincome | int | Stores max salary of a salary slab |
| taxper | double | Tax (%) for a salary slab |

# Job 2 : Financial Analysis

- Final Output
  - Task1
    - (1728.2616350000017)
  - **Task2**
    - (adult,1828.8573029045647)
    - (elderly,1822.1314705882355)
    - (infants,1632.2484130982368)
    - (Teenager,1758.8825362318848)
    - (middle-aged,1671.9536820083677)
    - (senior citizen,1662.542173913042)
  - **Task3**
    - ( Male,1792.9552289669857)
    - ( Female,1671.0068897266733)

# Job 3 : Social Analysis

- Task 1 : Calculate :
  - Pension Amount to be added after x years
  - No. of Orphans for each category based on Parents Present
  - No. of Employable Female Citizens who are Widows or Divorced
- Supporting Tables Required :
  - Pension_Mapping Table:
  - Orphan_Mapping Table:

| Column Name | Column Type | Remarks |
|---|---|---|
| Pid | int | Primary Key |
| min_income | int | Stores min salary of a salary slab |
| max_income | int | Stores max salary of a salary slab |
| pension | int | Pension amount paid |

| Column Name | Column Type | Remarks |
|---|---|---|
| oid | int | Primary Key |
| parent_present | varchar | Status of Parents present |
| subsidy | int | Subsidy amount paid to orphan |

# Job 3 : Social Analysis

- Final Output
  - Task1
    - Total Pension Amount
  - Task2
    - **( Not in universe,1472)**
    - **( Father only present,17)**
    - **( Mother only present,133)**
    - **( Neither parent present,18)**
  - **Task2**
    - **( Widowed,75)**
    - **( Divorced,70)**

# Job 4 : Planning Analysis

- Task 1 : Calculate :
  - No. of Voters to get added in next X years
  - No. of Senior Citizen to get added in next X years
  - Sex Ratio
  - Citizen vs. Immigrants Ratio for all Employed

# Job 4 : Planning Analysis

- Final Output
  - Task1
  - Task2
  - Task3
    - **( Male,939)**
    - **( Female,1061)**
  - Task4
    - **(135,1865)**


Any
Questions


Thank
you