Indiana Oracle Users Group – Technology Day – May 4th, 2016

Introduction to Big Data

Ashokkumar Sivasankaran

ensono | OPERATE FOR TODAY. OPTIMIZE FOR TOMORROW.

About me

- Ashokkumar Sivasankaran (Ashok)
- Client Technical Lead/Database Architect, Ensono
- Twenty eight years in IT services
- OCE RAC Expert & OCP Database Administrator 7.3 to 11g
- ITIL V3 Foundation Certified
- Member IOUG, COUG and INOUG
- https://www.linkedin.com/in/ashokkumar-sivasankaran-53418828
- Twitter: @ashokkumarsivas

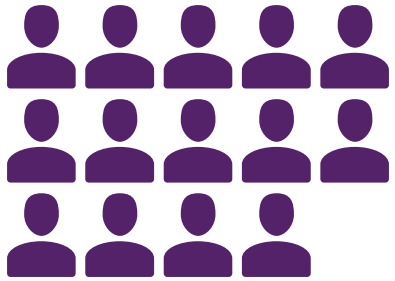
Fast Facts about Ensono

- Start-up with 46 years of experience managing complex infrastructures of the world's most successful companies
- Specialize in supporting mission-critical workloads
- 700 IT associates worldwide, majority in the U.S.
- Headquartered in greater Chicago, IL
- 9 North American and 2 international data centers
- Over $200 million in revenue
- Recognized by Gartner, InformationWeek and The Black Book of Outsourcing for service excellence and innovation

The Power in Numbers

MORE THAN
10,000 SERVERS
UNDER MANAGEMENT

AVAILABILITY –
UPTIME INSTITUTE
FIVE CONSECUTIVE YEARS
100%

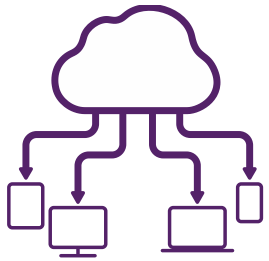


14 YEARS
AVERAGE TENURE
OF EMPLOYEES

46 YEARS
OF EXPERIENCE

10 PETABYTES
OF STORAGE
BEING MANAGED

OVER
61,000
MAINFRAME MIPS
MANAGED



MORE THAN
11,000
DATABASES
MANAGED

We are recognized by some of the leading experts in technology

Recognized by the Uptime Institute for achieving **100% AVAILABILITY IN DATA CENTERS** over the last three years

Ranked as "**#1 IN CUSTOMER SATISFACTION FOR IT OUTSOURCING**" by the Black Book of Outsourcing

Ranked "among the **TOP 3 MAINFRAME OUTSOURCING PROVIDERS** in North America and in the **TOP 10 DATA CENTER OUTSOURCERS**" by Gartner

InformationWeek 500 ranked "Ensono in the **TOP 3 FOR TECHNOLOGY INNOVATION** in the **BUSINESS SERVICES CATEGORY**"

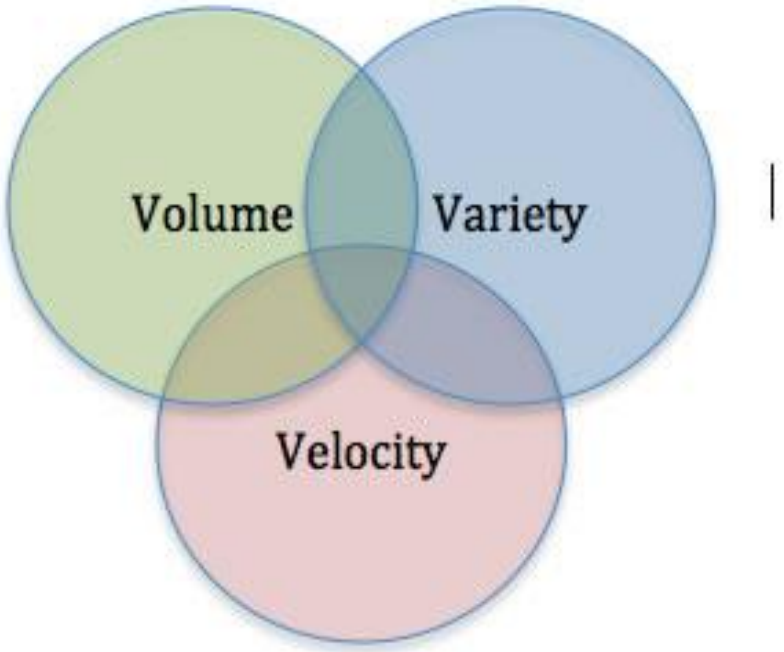
Data, Database, RDBMS and Big Data

- Data - Data are values of qualitative or quantitative variables, belonging to a set of items. It can be structured, un-structured and semi-structured data
- Database - A structured set of data held in a computer, especially one that is accessible in various ways.
- RDBMS - Relational Database Management System(RDBMS), invented by E. F. Codd at IBM in 1970. Data stored in normalized form and reduces data duplication. i.e., employee master table and monthly salary table
    - Major RDBMS are: Oracle, Sqlserver, DB2 and Sybase
- Big data - Superset of all types of data – mainly unstructured

What is Big Data?

- Collection of data sets – large and complex(Big data is a superset. RDBMS/structured data can be part of it)
- Difficult to process through traditional rdbms and database application.
- Higher cost to process large data volume using traditional RDBMS
- Data can be
  - Structured - rows and columns
  - Unstructured – web logs/click stream/binary/images
  - Semi structured - document
- Big data deals mostly with unstructured data
- NoSql database(commonly interpreted as “not only SQL”)
- NoSql databases generally don’t use SQL for DML

Big Data



Why we need it?

- Better understanding about customer preference, market trends
- **Predict**, plan and take proactive actions
- Converting challenges/changes to growth opportunities
- Thinking ahead of the curve, product issue postings at social network about a product(i.e., cell phone)
- Out of box thinking from pattern, correlation, outliers
- Nowcasting vs Forecasting
    - Nowcasting methods based on social media content (Google, Twitter) have been developed to estimate hidden quantities such as the 'mood' of a population or the presence of a flu epidemic.

Big Data use cases

- Content Optimization and Engagement Modeling
- Loyalty, Promotion Analysis and Targeting
- Usage Analysis and Mediation
- Entity Surveillance and Signal Monitoring
- Network Analysis and Sessionization
- Recommendations and Modeling
- Time series Analysis, Mapping and Modeling
- Fraud Analysis, Reconciliation and Risk

Case studies: http://searchcio.techtarget.com/opinion/Ten-big-data-case-studies-in-a-nutshell

How to approach?

- Start with questions – What we want to know? How do we want it? What will be the benefit?
- Develop Big Data strategy
- Connect the Stakeholders
- Identify data, data source, extract, clean, aggregate and analyze processes
- Identify right tools
- Architect a pilot project to meet the above requirement
- Don’t allow scope creep
- Validate results at each stage

Implementation

- Select software according to data type and processing requirement
- Use commodity low cost servers or virtual servers or **cloud infrastructure**
- Dynamic distribution and parallelism
- Efficient Map-Reduce, aggregation
- Parallel, in-memory, minimize network communication
- Fault tolerance and automatic recovery
- Moving computation to the data(moving data is expensive)
- Data processed in chunks in multiple nodes(distributed computing)
- High availability and scalability

Hadoop

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

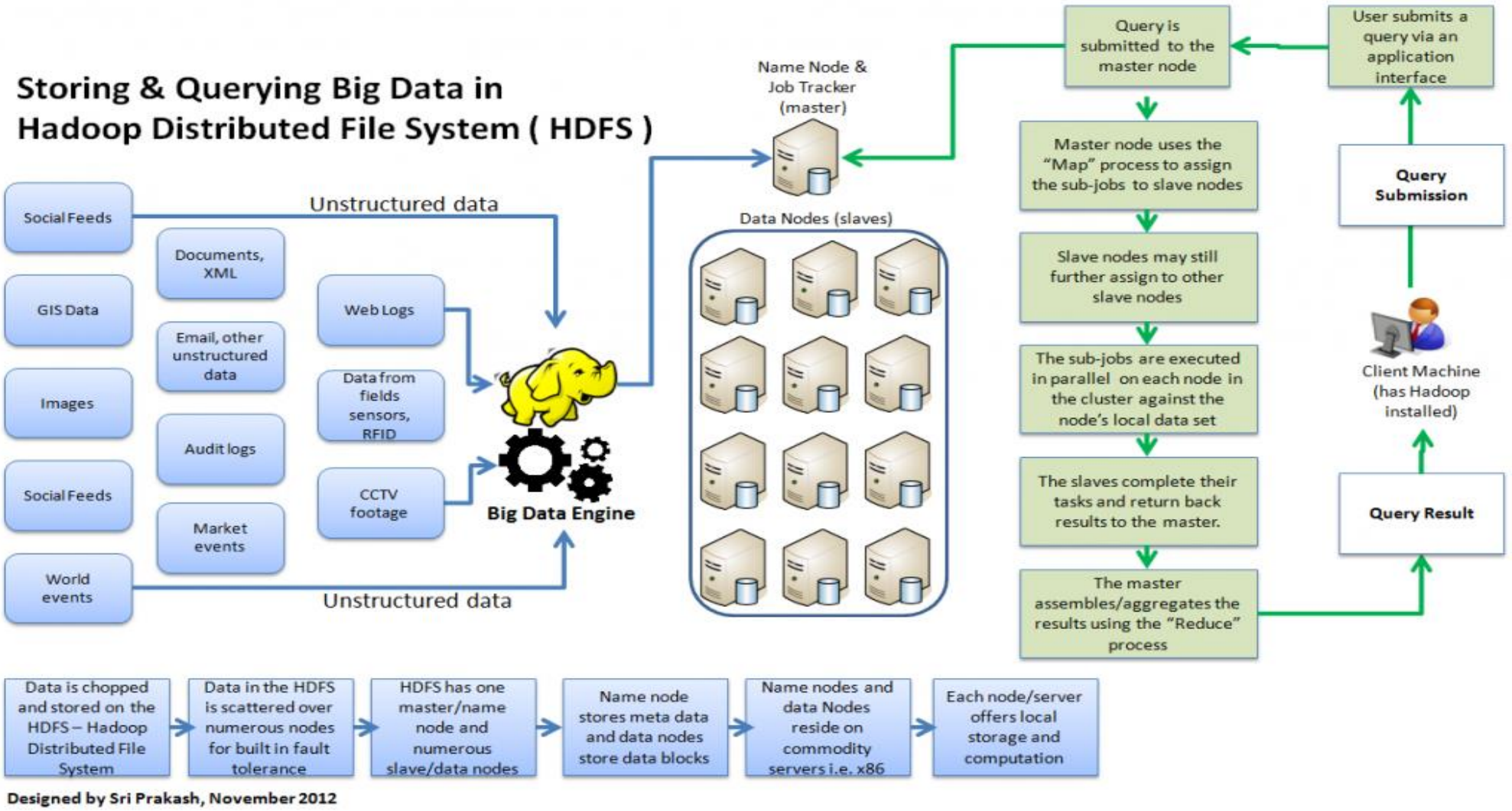
Source: Wikipedia

Hadoop is for offline batch processing and not for online and transaction process(OLTP)

Hadoop consists of two primary components:

- HDFS – Distributed file system for storage
- MapReduce – Parallel processing framework

Hadoop Architecture



NoSql Databases

- MongoDb (document, Native JSON support,..)
- Redis (Stores all data in memory, good for real-time analytics)
- Cassendra (developed at Facebook, key-value store column based)
- CouchDB (JSON access over HTTP, best for web access)
- Hbase (integrated with Hadoop architecture, Tabular/columnar storage, non-relational, distributed database, best for complex computing jobs)

Hadoop Distributions, other tools and Software

Hadoop Distributions

- Amazon Web Services
- Cloudera
- Hortonworks
- MapR

Other tools

- ZooKeeper – Coordination framework
- Hive – SQL like interface for querying
- Pig – High level scripting language for batch processing and ETL

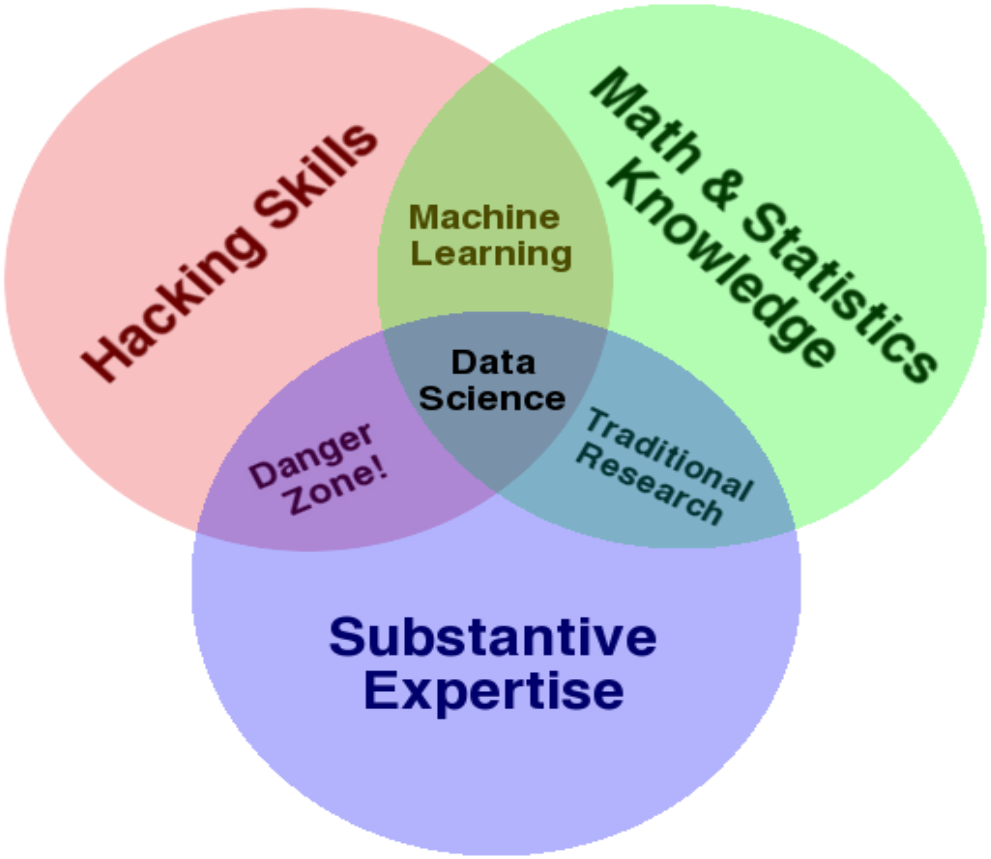
Hadoop Distributions, other tools and Software

- Flume – Distributed service for collecting, aggregating, and moving large data
- Sqoop – Bulk data transfer
- Mahout – Data Mining utility using machine learning
- Hue – Web console for Hadoop
- Oozie – Management workflow and job scheduling
- Storm – Distributed computation framework
- Kaffka – open source message broker
- Impala – Cloudera Impala is massively parallel processing SQL query engine for data stored in Hadoop cluster
- YARN – Resource management system

BIG data workflow

Support Area	Resource	Skills required
Developing and Architecting Solution	Big Data Architect	Understanding Business need for Big data solution, data source, moving data, processing methods, infrastructure need and developing cost effective flexible solution
Hardware, Storage and OS support	Hardware, Storage, OS and VM Admin	Linux admin experience to build physical or virtual linux servers to create Big data cluster
Network and Security	Network and Security Admin	Providing secured network for the Linux servers
Resource and Availability Monitoring	Monitoring Admin	Monitor servers and key processes
Backup and restore	Backup Admin	Backup and restore identified file systems
Big data admin(Hadoop or other software)	Hadoop/No Sql(MongoDB/CouchDB/..) Admin	Administering Big data cluster
Scheduling Data collection, processing and sending data jobs	Job scheduler/Middleware Admin	Operation, Development and Middleware support
Data collection and processing	Developer	Strong Java /Python programming and Shell scripting experience
Analytics and Data visualization	Data Analyst/ Data scientist	Analyzing, pattern discovery and predicting

Data Science



Machine Learning

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can teach themselves to grow and **change when exposed to new data**.

Source: Techtarget

Example:

- Spam filter

Data Analytics and Visualization

Data analytics is the science of examining data using quantitative and qualitative techniques, finding relationship, pattern, outliers and predict.

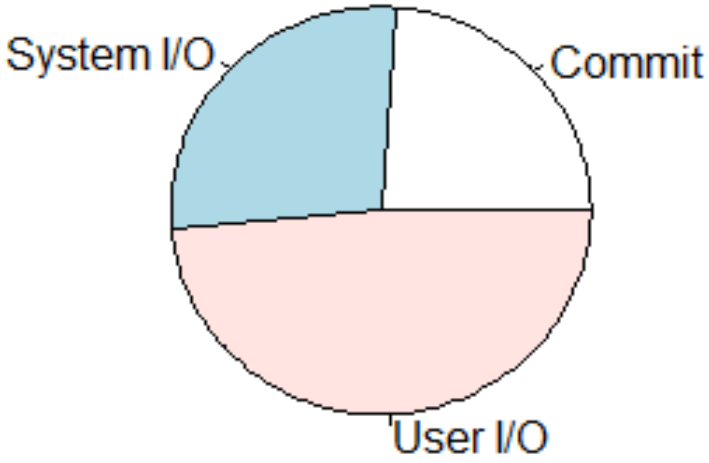
Data visualization is the presentation data in pictorial format. It helps to understand data, pattern, impacts and make decision.

R programming

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing.

```
> table(awr_data$WAIT_CLASS)
   Commit System I/O   User I/O
     1197       1378       2425

> pie(table(awr_data$WAIT_CLASS))
```



Oracle solutions to Big Data

- Oracle Big Data Cloud Service
- Oracle Big Data Machine
- Oracle Big Data Sql
- Oracle Advanced Analytics
- Oracle Data integrator for Big data

Oracle is providing Oracle Big Data Lite virtual machine for testing and educational purpose

Questions ?

Thank you!