Kathmandu Apache Spark Meetup

INTRODUCTION TO APACHE SPARK

Lightening Fast Cluster Computing

Saturday, 2015-12-19 @ Hotel Shanker, Kathmandu.

Hosted by RoyalePi

AGENDA

- About Kathmandu Apache Spark Meetup
- About RoyalePi
- Q&A by Databricks
- Big Data Overview
- Apache Spark Introduction
- Hello from Paris by Ludwine
- Break for Snacks - 12:00 - 12:30
- Hands-on
- Crime against Women in India case study using R & Python by Aakash Bikram Chand
- Quiz

KATHMANDU APACHE SPARK MEETUP

- It's a community of professionals and enthusiasts
- We are **51** members as of today
- Once a quarter - can vary
- Individual - share, learn and network
- Companies - promote, attract talents
- Hands-on better than just the presentation
- Presentation and technical materials to be shared
- Speakers - local, remote (US and other countries)

ABOUT ME

Software Engineer (Data & Analytics) @ Guidewire Software, San Francisco Bay Area

Twitter @geechand

LinkedIn chandganesh

EDUCATION

- School
  - DPS, Doti
  - LRI, Kathmandu
- College
  - AMC PU College, Bangalore
  - BVB College of Engg. and Tech, Hubli, Karnataka
- Others
  - Continuing Studies - Stanford University and UCSC Santa Cruz

PROFESSIONAL EXPERIENCE

- 11 years
- Past Companies: SunGard, Accenture, Hitachi, American National Insurance Company
- Places: Bangalore, Mumbai, Yokohama, Houston, San Francisco

ABOUT ROYALEPI

WHO WE ARE?

- Kathmandu based Startup, founded in 2013
- Development Center - Kathmandu and Dhangadhi
- We're in it to insπre
- 4 (+1) members

WHAT WE DO?

- Build Software Products - Transactional and Data Analytics
- Training for Data Scientist and Data Engineer
  - Big Data
  - Hadoop
  - Python
  - R
  - Spark
  - Scala
  - Machine Learning

WHAT'S NEXT?

- Data Driven Analytic applications - Work-in-progress
- Why Spark?

AKALICO OVERVIEW

OVERVIEW AND Q&A BY DATABRICKS

- Denny Lee
- Jason Pohl

BIG DATA OVERVIEW

WHAT IS BIG DATA?

It depends...

Google & Oxford

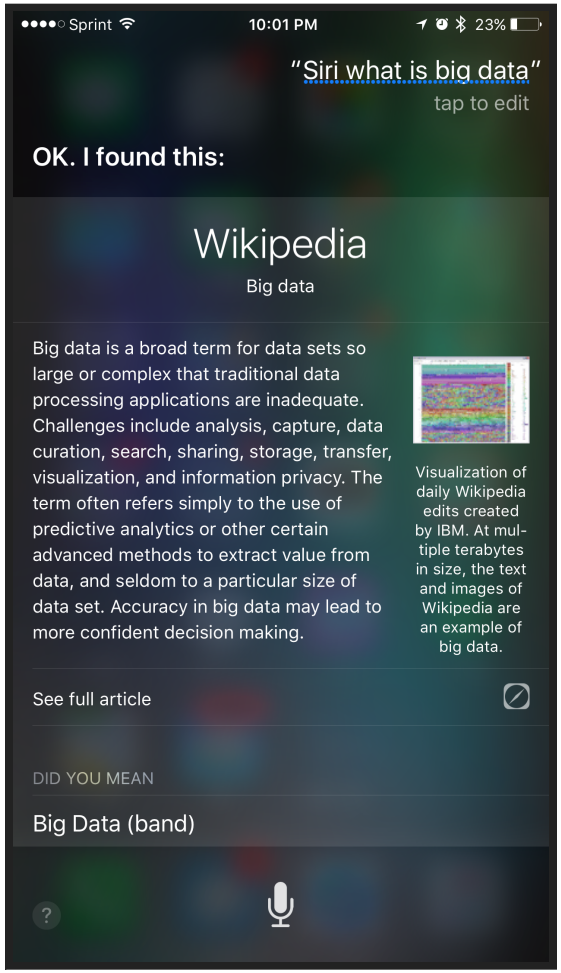
big da·ta

noun COMPUTING
noun: **big data**

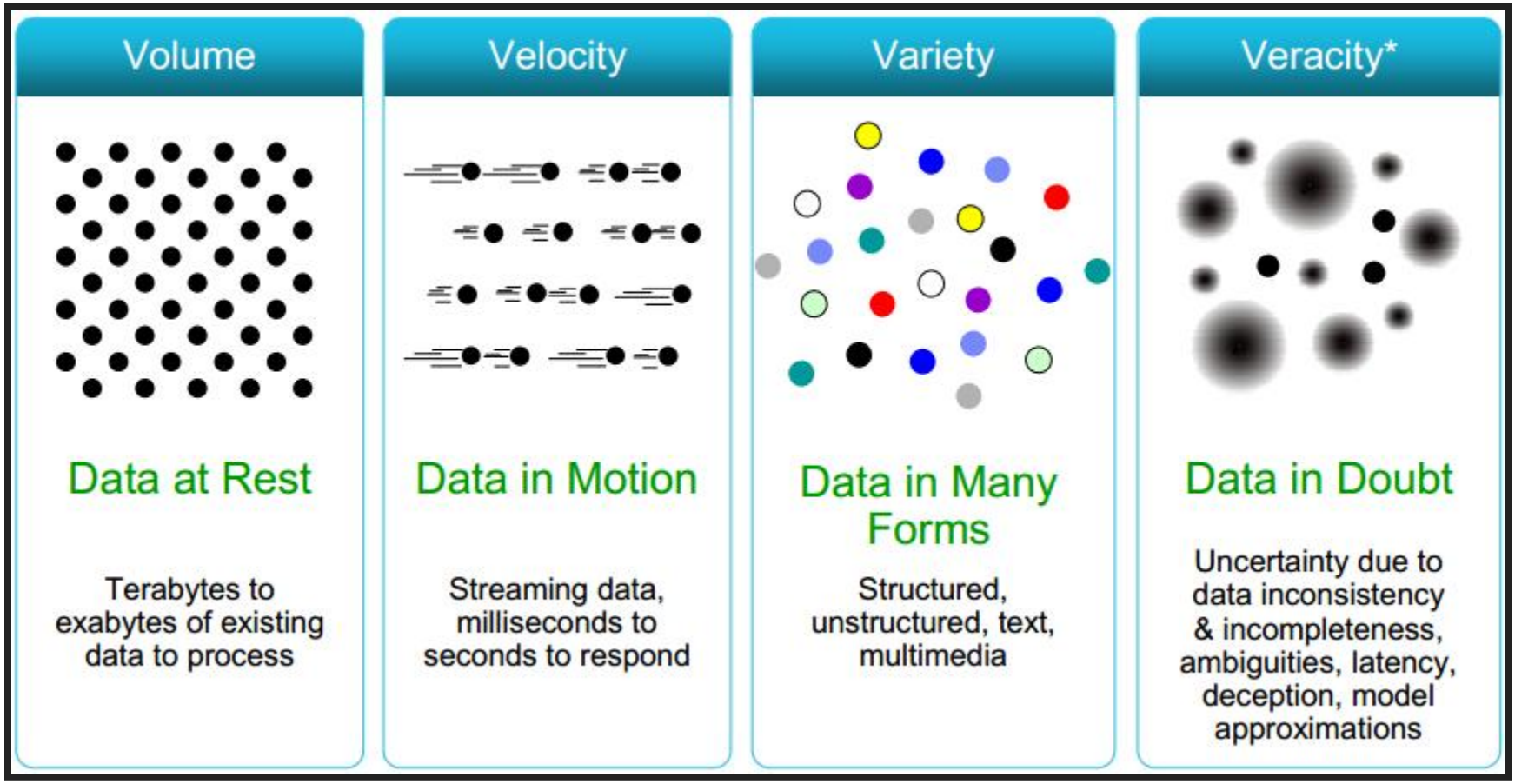
extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions.
"much IT investment is going towards managing and maintaining big data"

"Big data is a broad term for data sets so ***large or complex*** *that traditional data processing applications are inadequate."*
- wikipedia

As per **Siri...**


Sprint
10:01 PM
23%
"Siri what is big data"
tap to edit
OK. I found this:
Wikipedia
Big data
Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making.
Visualization of daily Wikipedia edits created by IBM. At multiple terabytes in size, the text and images of Wikipedia are an example of big data.
See full article
DID YOU MEAN
Big Data (band)

The Four Vs...



source:http://www.datasciencecentral.com/profiles/blogs/data-veracity

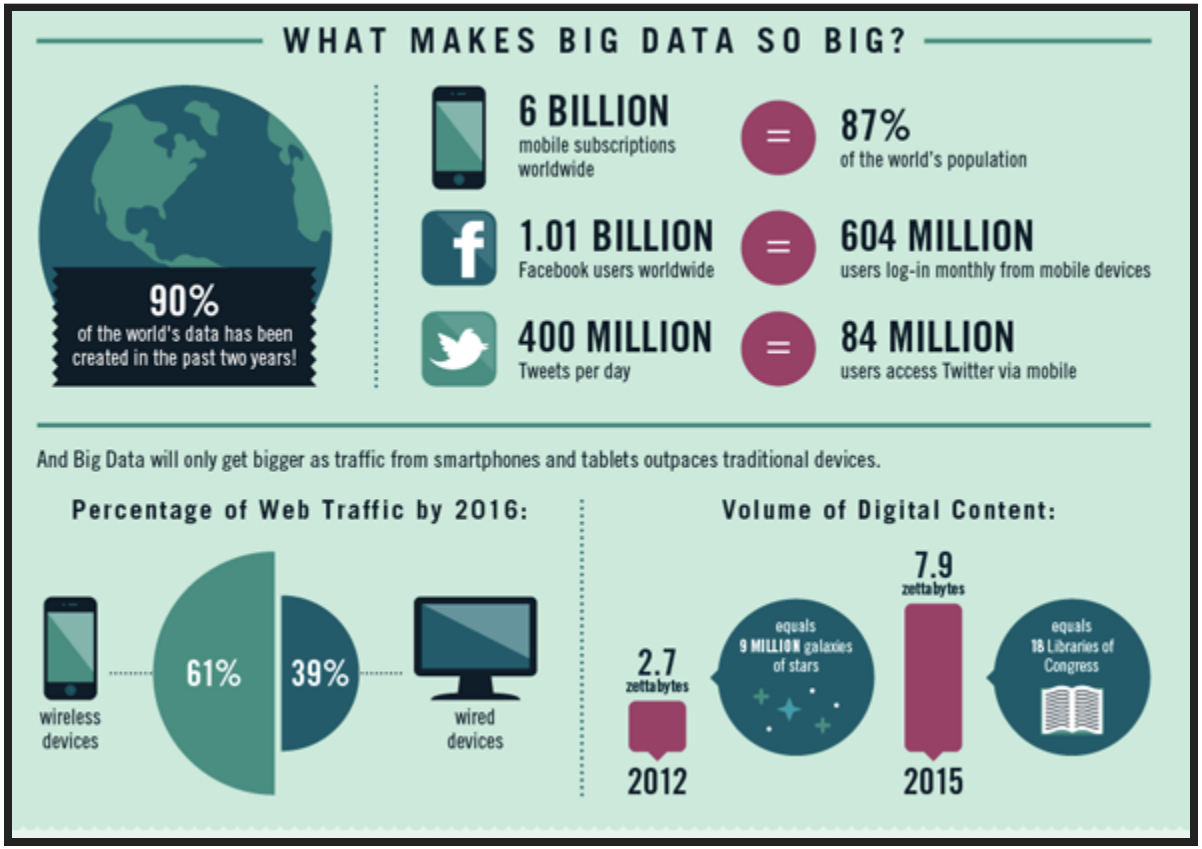
Well, there is the 5th "V" - Value.

- Most Important V of Big Data
- Cost & Benefits

Some use cases

- Understanding and Targeting Customers - amazon.com
- Understanding and Optimizing Business Processes - Talent Acquisition
- Personal Quantification and Performance Optimization - Fitness Apps
- Improving Healthcare service - Disease Prevention and Control
- Improving Sports Performance - Sports Science, MoneyBall
- Improving customer Service - Guest Analytics for Hotels
- Improving Security and Law Enforcement - Predictive

INTERESTING FACTS ABOUT BIG DATA



source: https://mydigitaleyeshadow.wordpress.com/

The NSA is thought to analyze **1.6%** of all global traffic - around **30 petabytes**(30 million gigabytes) **every day!**

Around **100 hours** of videos are uploaded to YouTube **every minute** and it would take you around **15** years to watch every video uploaded by users in one day.

If you burned all of the data created in just **one day onto DVDs**, you could stack them on top of each other and reach the moon - **twice**.

The big data industry is expected to grow from **US$20.2 billion** in 2013 to about **US$54.3 billion**.

1.9 million IT jobs will be created in the US by 2015 to carry out big data projects. Each of those will be supported by 3 new jobs created outside of IT - meaning a total of **6 million** new jobs .

*"There were **5 exabytes** of information created between the dawn of civilization through 2003, but that much information is now created **every 2 days**"*

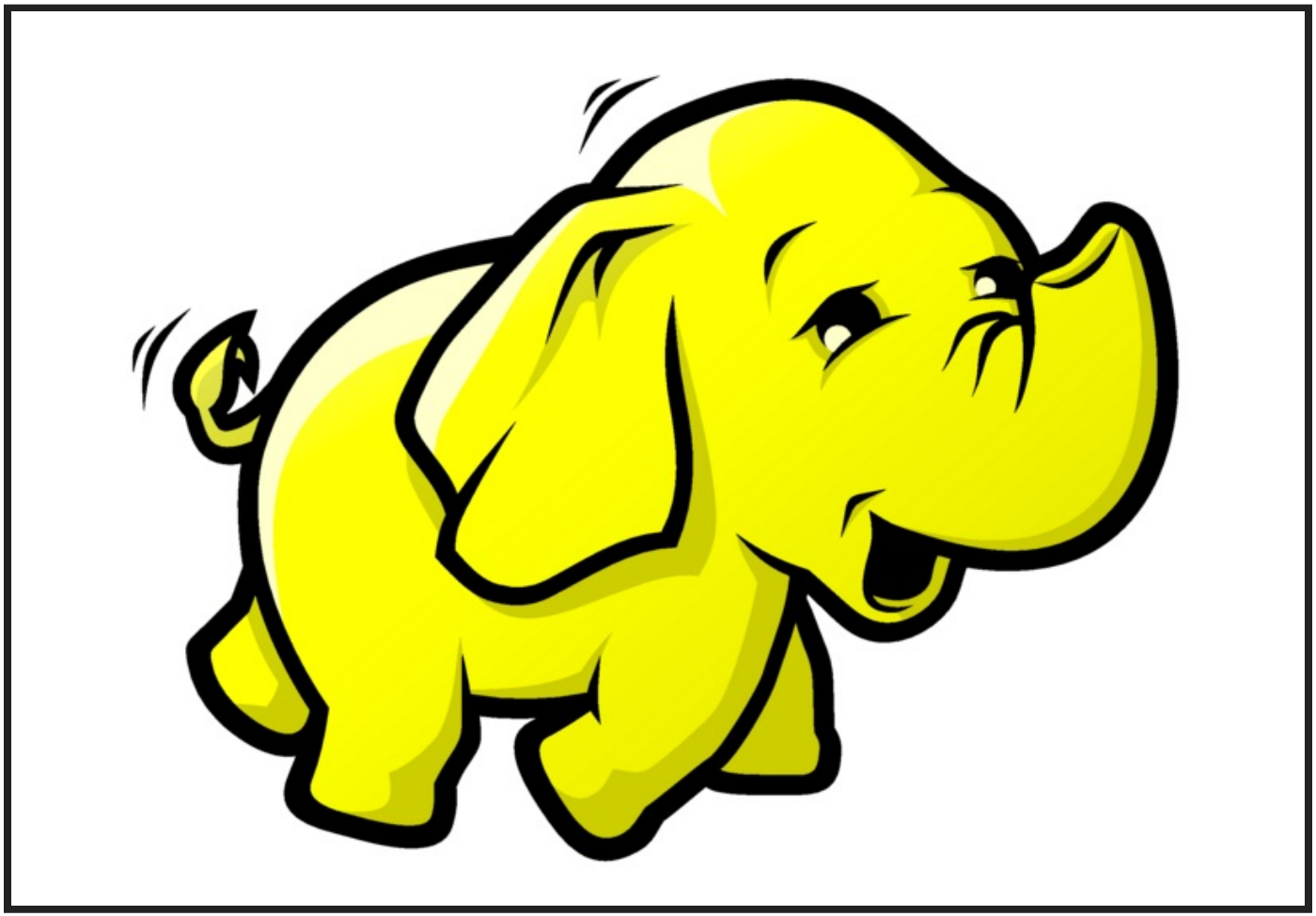
- Eric Schmidt, Google in 2010

---

*"Information is the **oil of the 21st century**, and analytics is the combustion engine."*
Peter Sondergaard, Gartner Research

No surprise - Facebook, Gmail, LinkedIn are all free!

HADOOP



WHAT IS HADOOP?

The Blind Men and the Elephant...

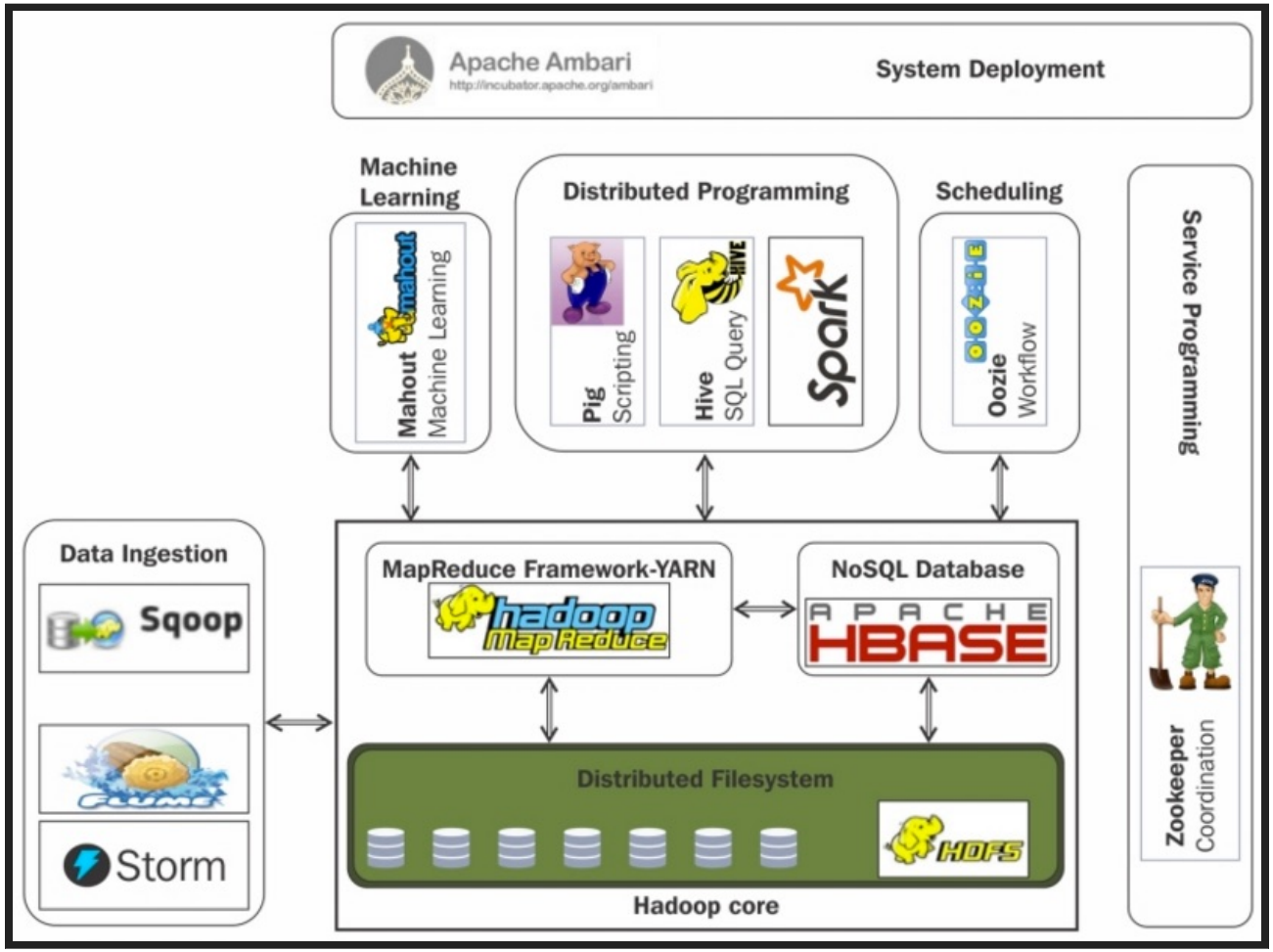
"After Hadoop finishes filtering the data, the place you want to put that data is in Oracle Database."

Larry Ellison (2011)

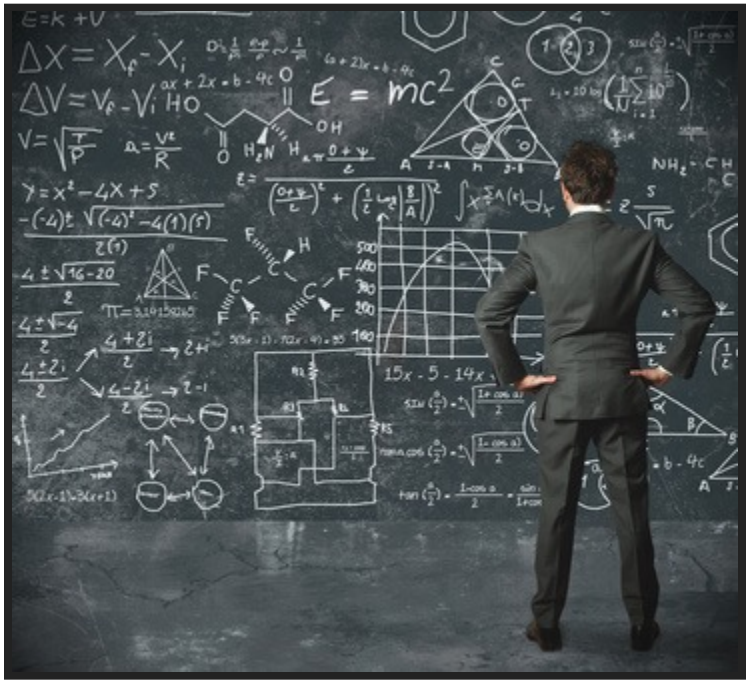
HADOOP IS AN OPEN-SOURCE FRAMEWORK FOR BIG DATA

- Hadoop was created by Doug Cutting and Mike Cafarella in 2005. Inspired by Google's white papers
- Hadoop is distributed, reliable and fault tolerant
- Horizontal scalability from single computer to thousands of cluster nodes.
- MapReduce White Paper by Jorrrey Dean and Sanjay Ghemawat - http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf
- The Four modules - Hadoop Common, HDFS, YARN & MapReduce

Hadoop Ecosystem



https://hadoopecosystemtable.github.io/



It's Complicated, I know...

APACHE SPARK INTRODUCTION

WHAT IS APACHE SPARK?

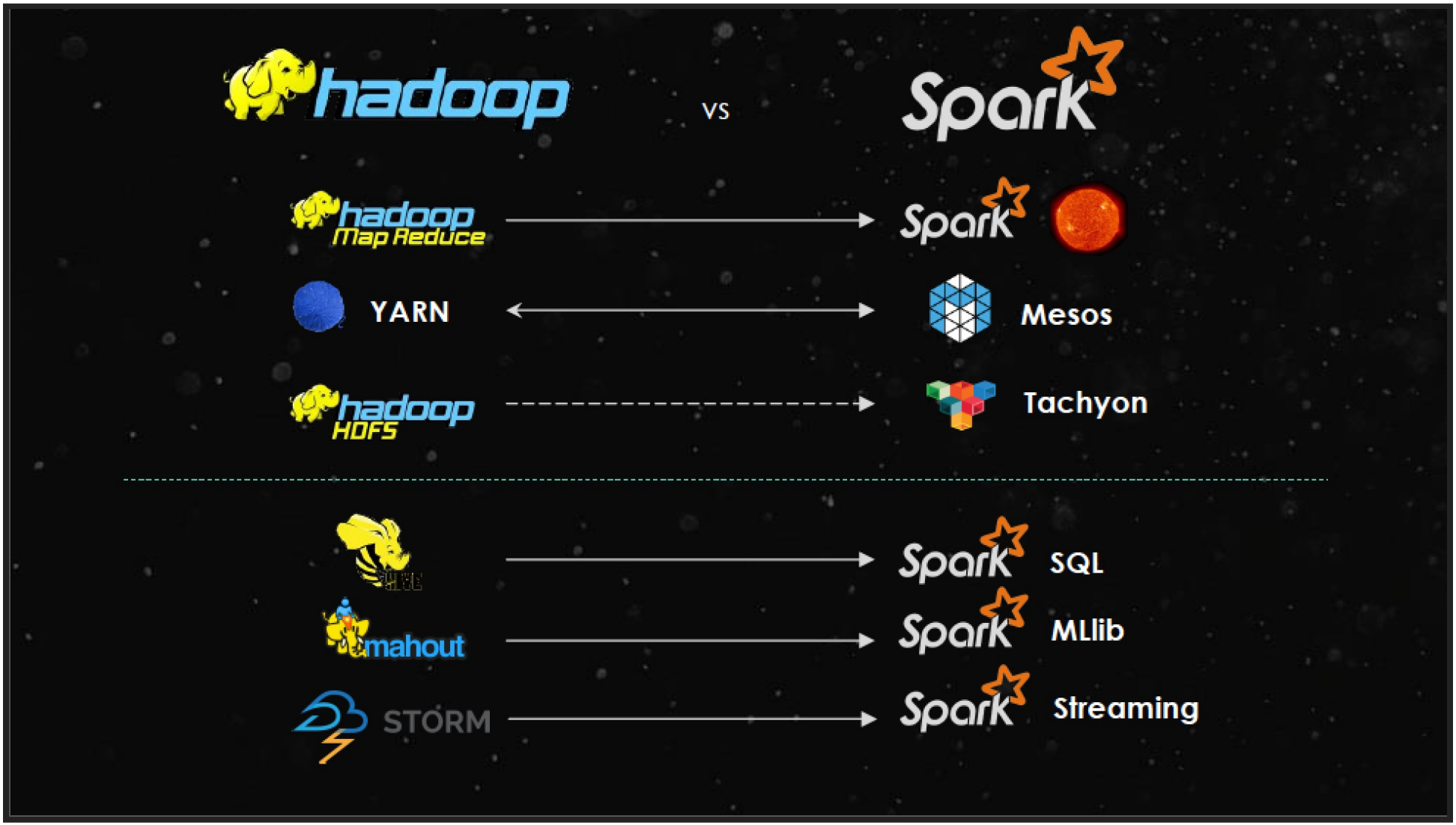
Apache Spark is a cluster computing platform designed to be fast and general-purpose

- **Speed** - ~100x times faster than MapReduce
- **Generality** - Batch, iterative, interactive, stream
- **Accessibility** - APIs in Java, Scala, Python, R
- **Spark can run in Hadoop clusters**

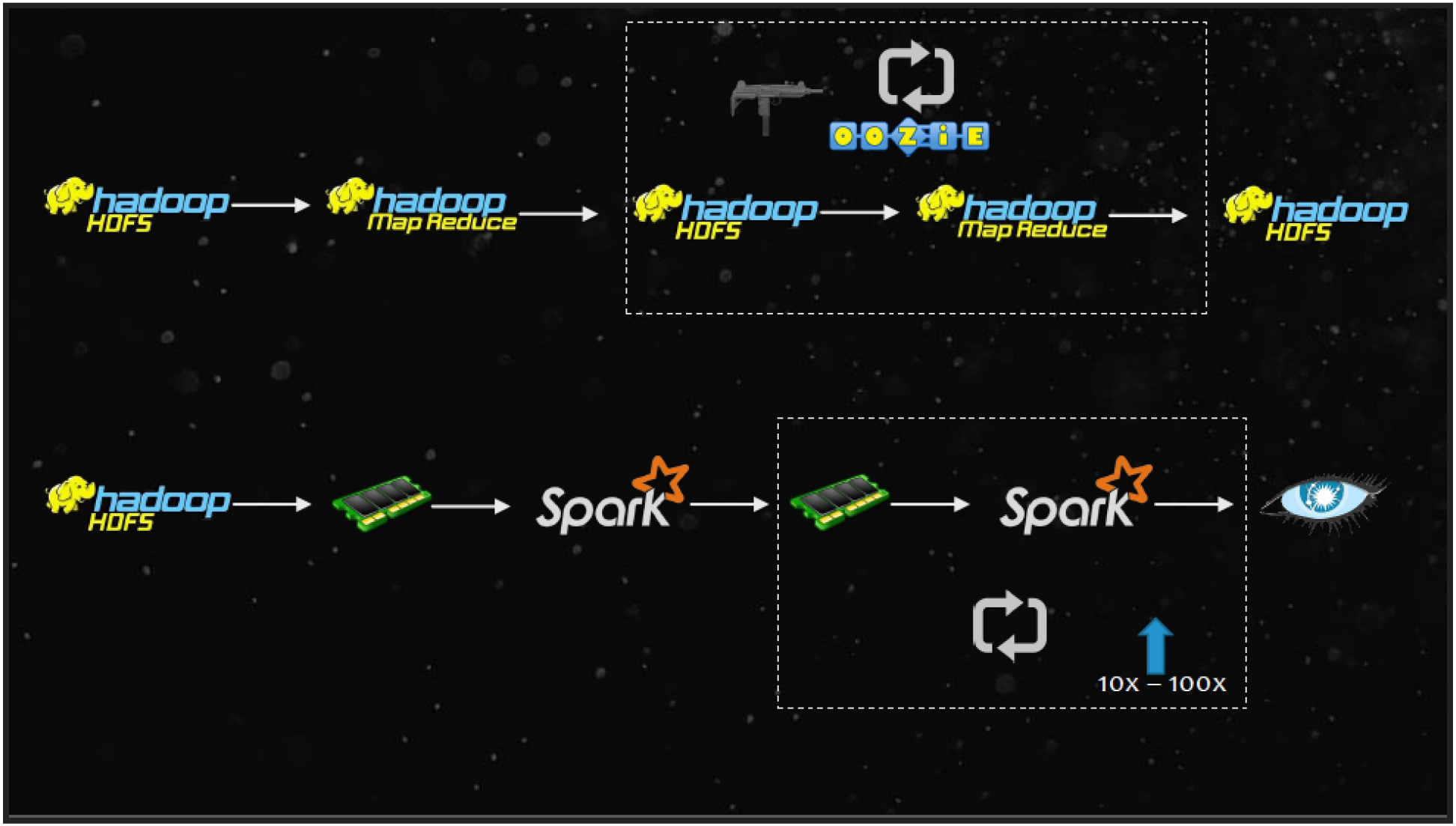
HISTORY OF SPARK

- Created in 2009 at UC Berkeley AMPLab
- Mostly Written in Scala
- Open Sourced in 2010
- Spark Paper published by Matei Zaharia, Mosharaf Chowdhury, Michael J. Franklin, Scott Shenker and Ion Stoica
- More than 400 developers from more than 100 organizations
- It is currently the most active Apache project

SPARK VS HADOOP - TOOLS



SPARK VS HADOOP - DATA PROCESSING



COMMERCIAL PACKAGING AND DISTRIBUTIONS

```
/** Scala REPL **/

    $ cd $SPARK_HOME
    $ bin/spark-shell --master "local[*]" --driver-memory 2G
```

- Spark context available as **sc**.
- SQL context available as **sqlContext**.

RDD(RESILIENT DISTRIBUTED DATASET)

- Main Abstraction in Spark
- A distributed memory abstraction that lets programmers perform in-memory computations on large clusters in a fault-tolerant manner
- Two ways to create RDDs:
  - Parallelizing an existing collection in your driver program
  - Referencing a dataset in an external storage system, such as a shared filesystem, HDFS
- White Papers
  - June 2010 - http://www.cs.berkeley.edu/~matei/papers/2010/hotcloud_spark.pdf
  - April 2012 - http://www.cs.berkeley.edu/~matei/papers/2012/nsdi_spark.pdf

CREATING RDD USING AN EXISTING COLLECTION

```
# Parallelize in Python
wordsRDD = sc.parallelize([“fish", “cats“, “dogs”])
```

```
// Parallelize in Scala
val wordsRDD= sc.parallelize(List("fish", "cats", "dogs"))
```

```
// Parallelize in Java
JavaRDD<String> wordsRDD = sc.parallelize(Arrays.asList(“fish", “cats“, “dogs”));
```

PARALLELIZE

- Take an existing in-memory collection and pass it to SparkContext’s parallelize method
- Not generally used outside of prototyping and testing since it requires entire dataset in memory on one machine

CREATING RDD USING EXTERNAL DATASETS

READ FROM TEXT FILE

```python
# Read a local txt file in Python
linesRDD = sc.textFile("/path/to/README.md")
```

```scala
// Read a local txt file in Scala
val linesRDD = sc.textFile("/path/to/README.md")
```

```java
// Read a local txt file in Java
JavaRDD<String> lines = sc.textFile("/path/to/README.md");
```

- There are other methods to read data from HDFS, C*, S3, HBase, etc.

RDD OPERATIONS

- Transformations
  - map()
  - filter()
  - join()
- Actions
  - collect()
  - reduce()
  - count()
- All transformations in Spark are lazy
- Can persist an RDD in memory using the persist (or cache) method

```
/** Spark Scala REPL example **/

    val listOfNumber = (1 to 10).toList // List[Int]
    val numRDD = sc.parallelize(listOfNumber) // RDD[Int]
    numRDD.collect // Array[Int]
    numRDD.first // Int
    numRDD.take(2) // Int
    val evenNumRDD = numRDD.filter(_ % 2 == 0)
    val oddNumRDD = numRDD.filter(_ % 2 != 0)
    evenNumRDD.collect
    oddNumRDD.collect
    val unionRDD = evenNumRDD.union(oddNumRDD)
    unionRDD.collect
    val sortedUnionRDD = unionRDD.collect.sorted
    sortedUnionRDD.collect
/** end **/
```

SPARK SQL AND DATAFRAME

- **Spark SQL** is a Spark module for structured data processing
- A **DataFrame** is a distributed collection of data organized into named columns.
- registerAsTempTable()

```
/** Spark SQL and DataFrame **/

  // Create the DataFrame
  val df = sqlContext.read.json("examples/src/main/resources/people.j

  df.printSchema() // Print the schema in a tree format
  df.show() // Show the content of the DataFrame
  df.select("name").show() // Select only the "name" column

  df.registerTempTable("people") //register it as a table
  sqlContext.sql("select * from people").show
/** end **/
```

SPARK HANDS-ON

- **Pre-requisite** - Apache Spark is installed and working
- Installation steps avaialable at: http://github.com/ganeshchand/kathmandu-spark-meetup-intro-to-spark/notes/install-spark.md
- Data and Source code available at: https://github.com/ganeshchand/kathmandu-spark-meetup-intro-to-spark

CRIME AGAINST WOMEN CASE STUDY USING R AND PYTHON

QUIZ

- Who created Hadoop?
- What University was Apache Spark originally developed at?
- What year was Apache Spark made an open source technology?

THANK YOU!