

---

# SPARKLING BIG DATA

Ashis Parajuli  
Big Data Engineer Fusemachines

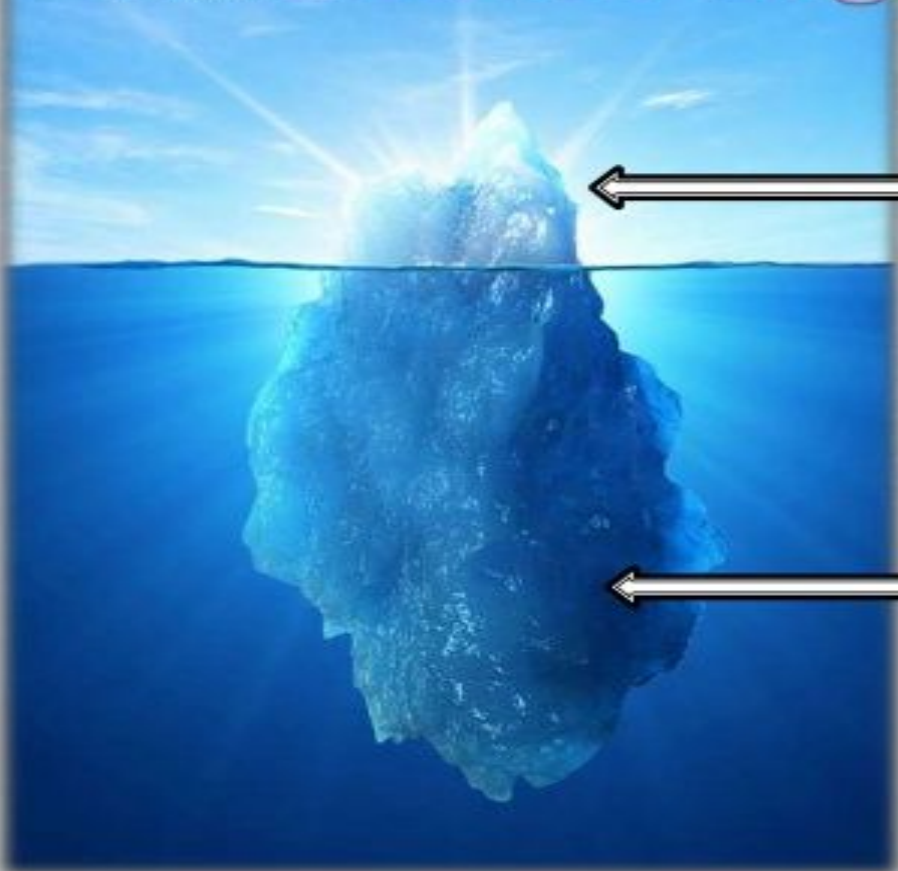


---

**Big Data ?**

---

# What does Big Data look like?



← What we know or see

← What's actually there

---

AI is an electricity then data is the coal/oil that drives it

---

# Agenda

---

- Big Data
  - Visualizations of data
  - Spark and/vs Hadoop
  - Big data “Hello World” program
  - Spark and it’s components
  - Demo on email classification using spark
-

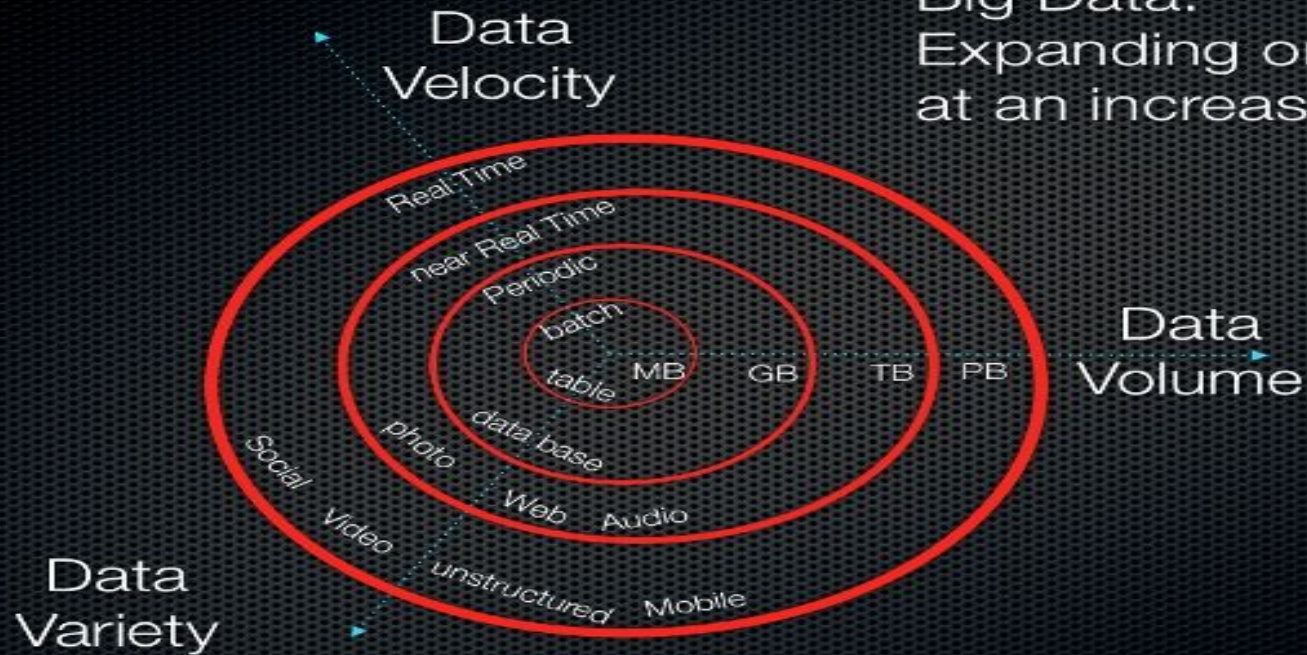


---

# 7v's

1. Volume
2. Velocity
3. Variety
4. Variability
5. Veracity
6. Visualization
7. Value.

Big Data:  
Expanding on 3 fronts  
at an increasing rate.





---

## Unstructured

I am Ashis Parajuli working at fusemachines.. I am Big Data Developer. I am currently staying at Koteshwor.

## Semi-Structured

```
{ name:"Ashis Parajuli",company: "Fusemachines",skills: "Big Data",Address:"Koteshwor"}
```

## Structured

Name	Company	Skills	Address
Ashis Parajuli	Fusemachines	Big Data	Koteshwor Kathmandu

---

## Scalability of NoSQL Database vs Traditional Relational Database

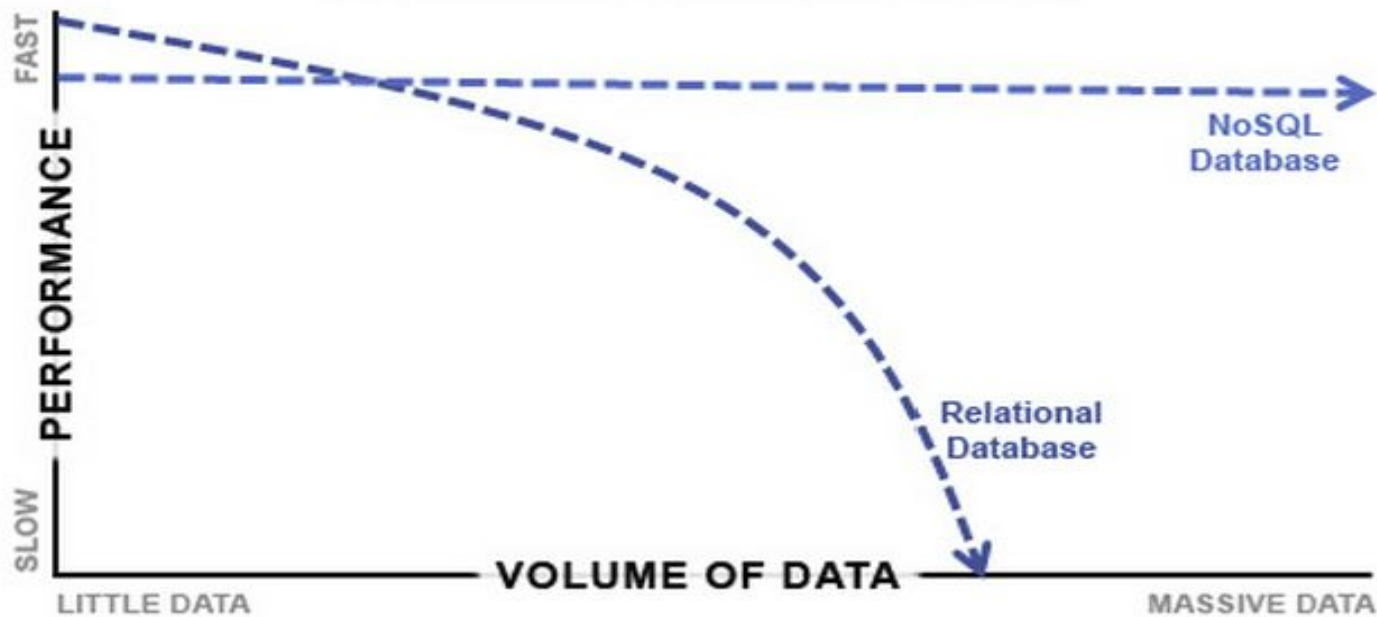


Image Credit: DataJobs.com

---

**Be Ready for the battle  
with Data**


---

---

# Data Visualization and Dashboard

---

# Data Visualizations

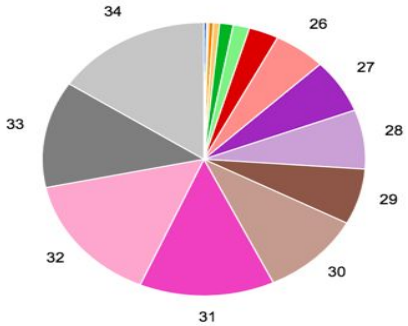
 Notebook ▼ Search your Notebooks   anonymous ▼

**Bank** ▶ ⌂ 📄 🗑️ 🔄 📄 ⚙️ 🔒 default ▼

### maxAge

 FINISHED ▶ ⌂ 📄 ⚙️  
  
📄 📊 📈 📉 📊 📈 👤 ▼ settings ▼  

- 19 20 21 22 23 24 25 26
- 27 28 29 30 31 32 33 34

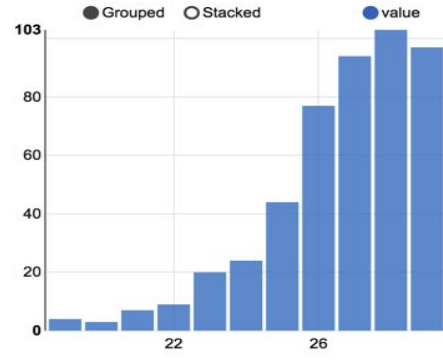


Took a few seconds. Last updated by anonymous at June 26 2016, 4:46:52 PM. (outdated)

### Under age < 35

 FINISHED ▶ ⌂ 📄 ⚙️  
  
📄 📊 📈 📉 📊 📈 👤 ▼ settings ▼  

● Grouped ○ Stacked ● value

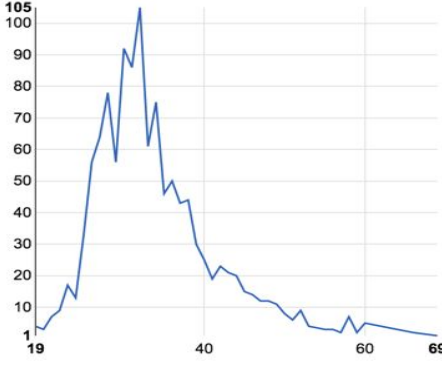


Took a few seconds. Last updated by anonymous at June 26 2016, 4:47:32 PM. (outdated)

### marital

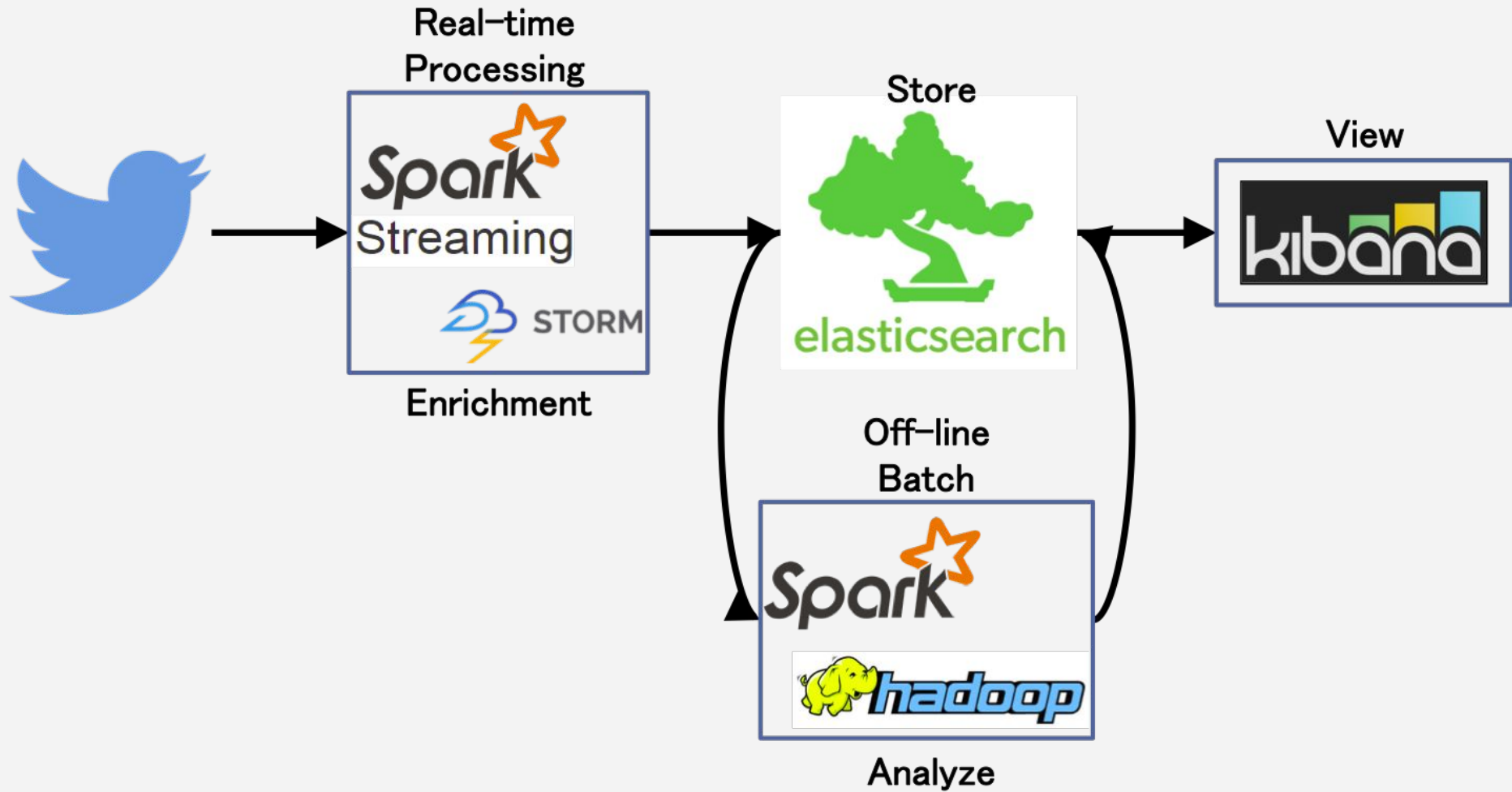
 FINISHED ▶ ⌂ 📄 ⚙️  
  
📄 📊 📈 📉 📊 📈 👤 ▼ settings ▼  

● value



Took a few seconds. Last updated by anonymous at June 26 2016, 4:47:36 PM. (outdated)

READY ▶ ⌂ 📄 ⚙️



# Applications for Big Data Analytics

Smarter Healthcare



Multi-channel sales



Finance



Log Analysis



Homeland Security



Traffic Control



Telecom



Search Quality



Manufacturing



Trading Analytics



Fraud and Risk



Retail: Churn, NBO



---

# Cloud Computing

---



---

# Public vs Private cloud

---

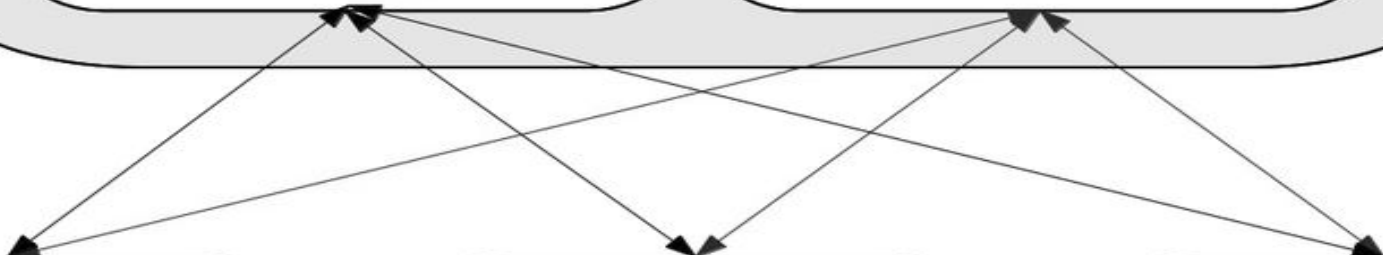
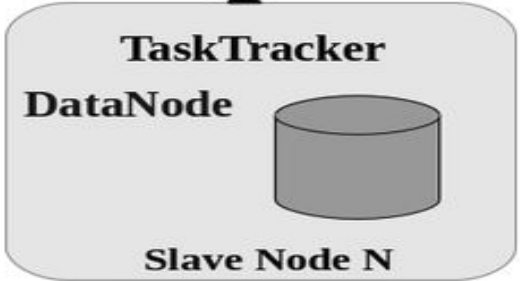
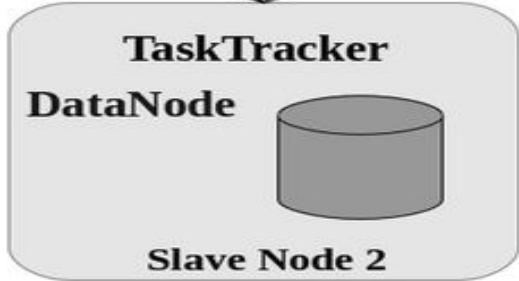
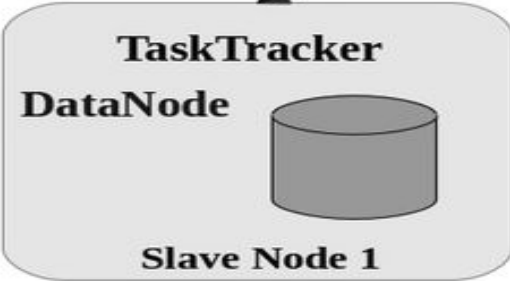
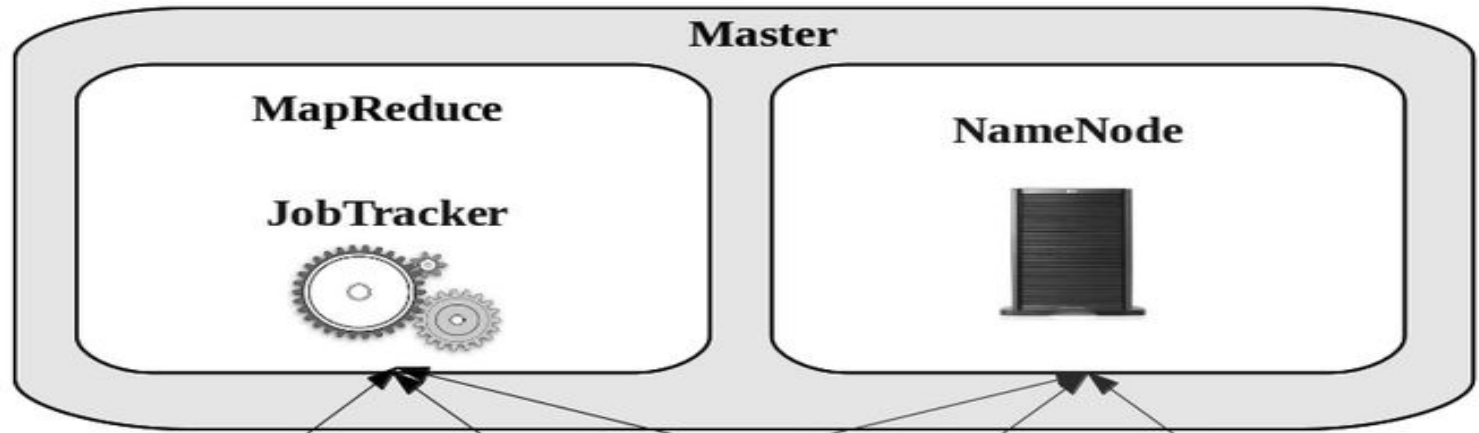
## Amazon Web Services(AWS)

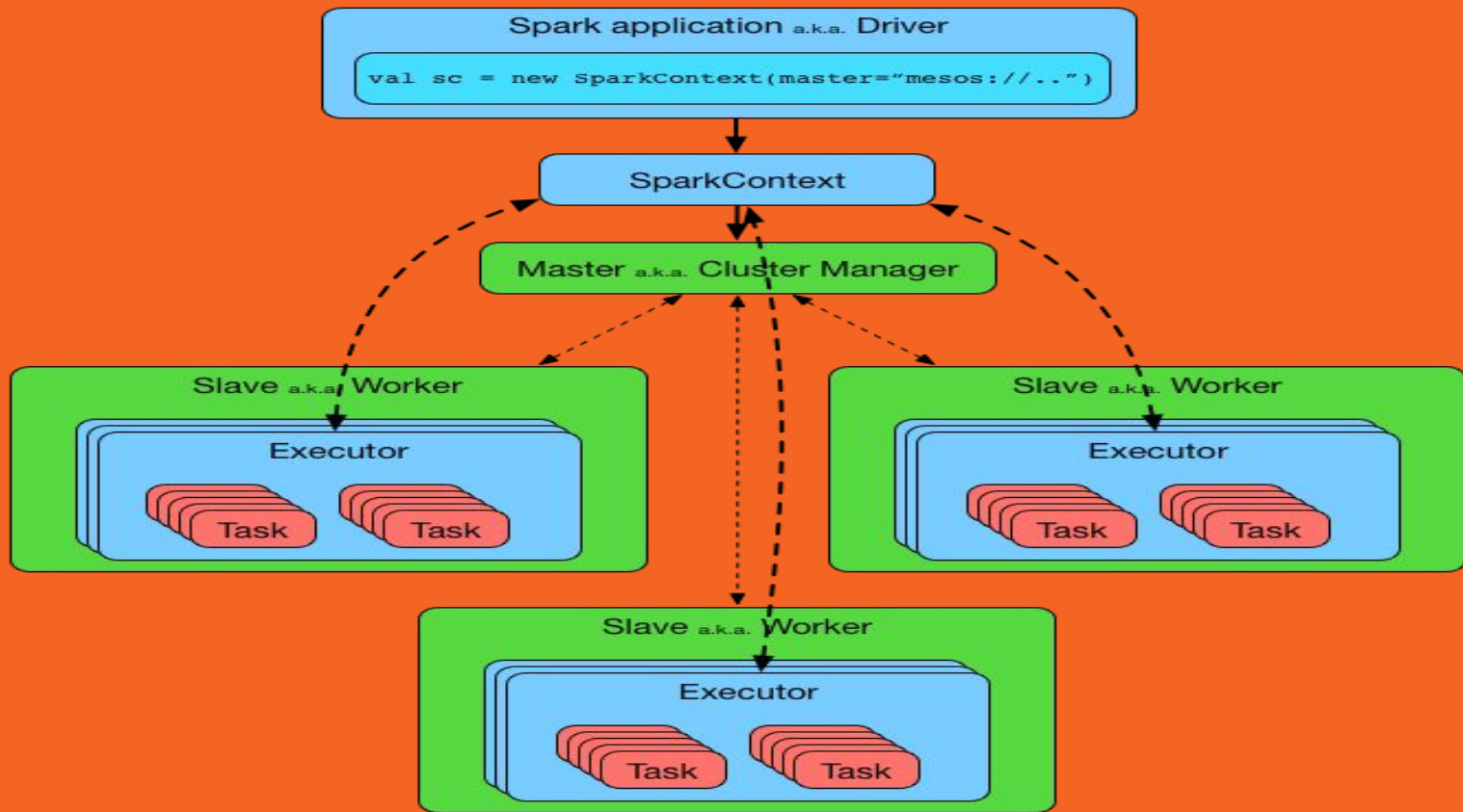
- Ec2
- S3

---

# Spark and Hadoop

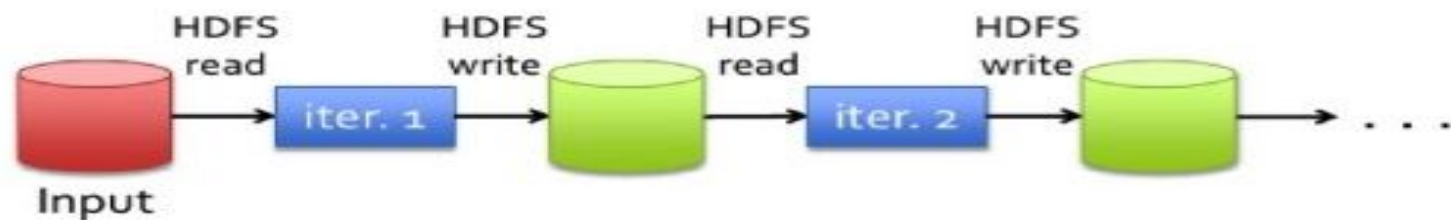
---





# HADOOP MAPREDUCE VS SPARK

---



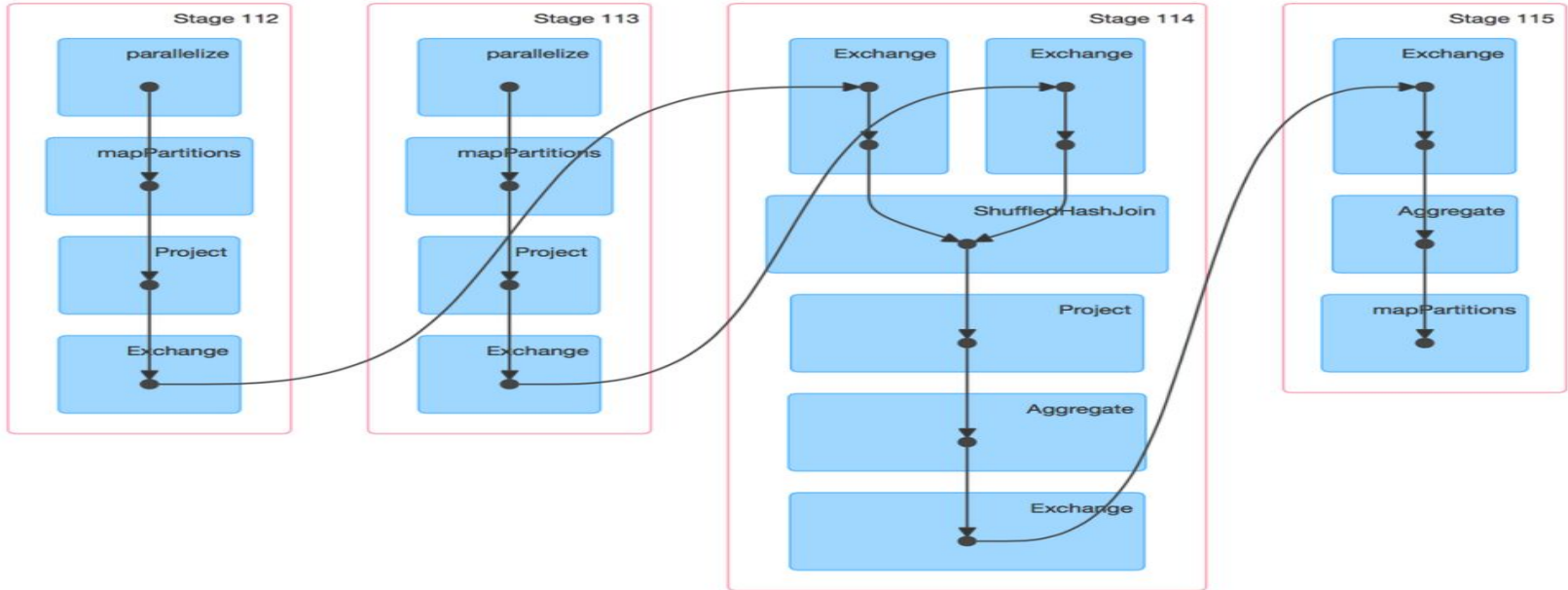
## Details for Job 8

Status: SUCCEEDED

Completed Stages: 4

▶ Event Timeline

▼ DAG Visualization



---

# Big Data Hello world Program(Word Count)

---



## Input

Hello world  
Hello visitor  
Visitor hello

## Mapper

Hello 1  
World 1

Hello 1  
Visitor 1

Visitor 1  
Hello 1

## Reducer

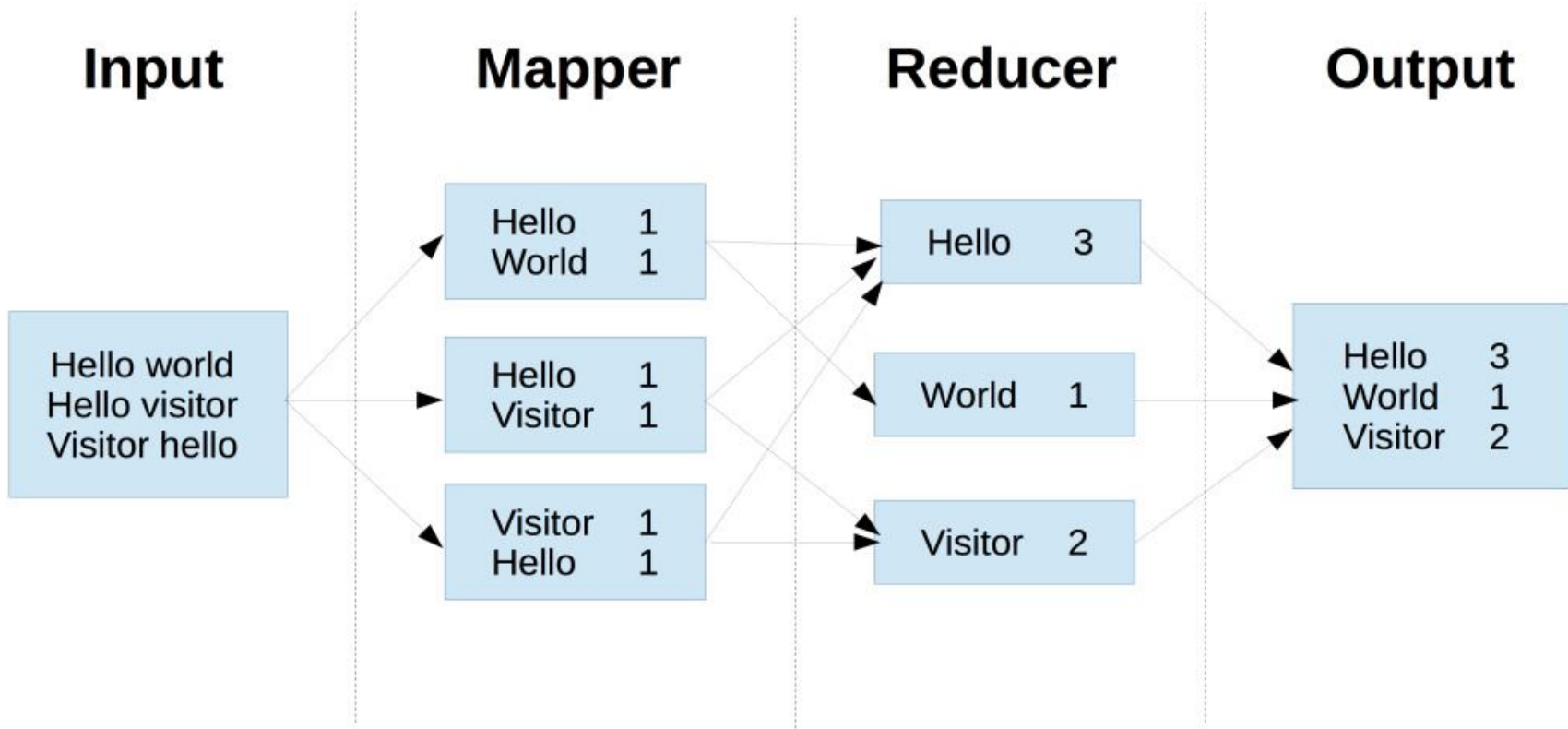
Hello 3

World 1

Visitor 2

## Output

Hello 3  
World 1  
Visitor 2



**RDD input**

```
w1 w2 w3  
w1 w1 w3  
w1 w3 w3
```

**flatMap(lambda x: x.split(' '))**

**RDD words**

```
[w1, w2, w3, w1, w1, w3, w1, w3, w3]
```

**map(lambda x: (x,1))**

**RDD words with  
initial counts**

```
[(w1,1), (w2,1), (w3,1), (w1,1), (w1,1), (w1,1),  
(w3,1), (w1,1), (w3,1), (w3,1)]
```

**reduceByKey(lambda x,y:x+y)**

**RDD words with  
final counts**

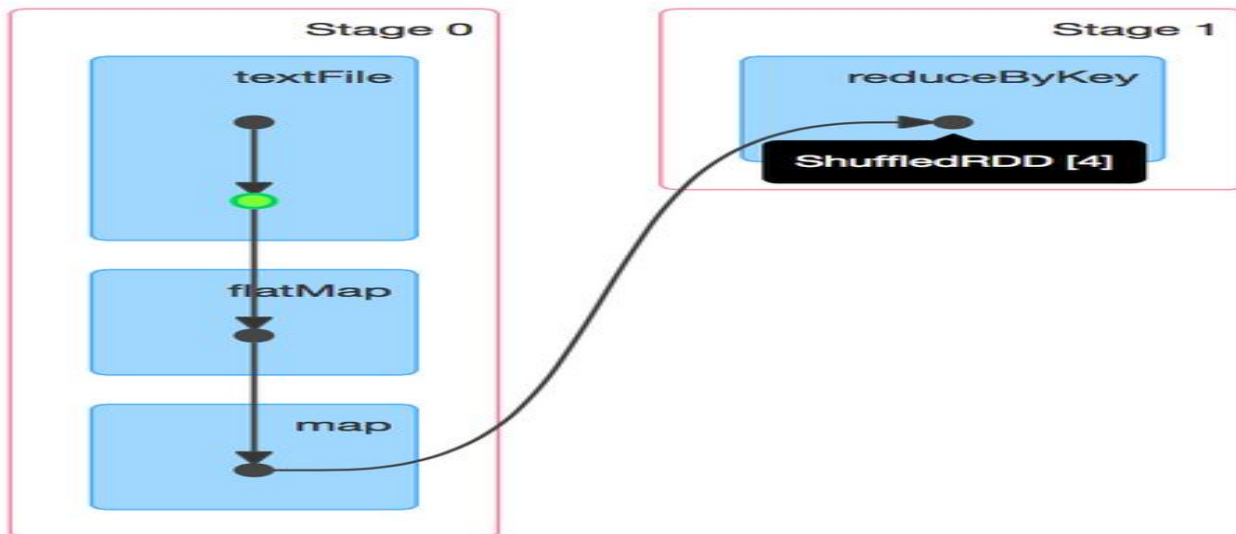
```
[(w1,4), (w2,1), (w3,4)]
```

## Details for Job 0

Status: SUCCEEDED

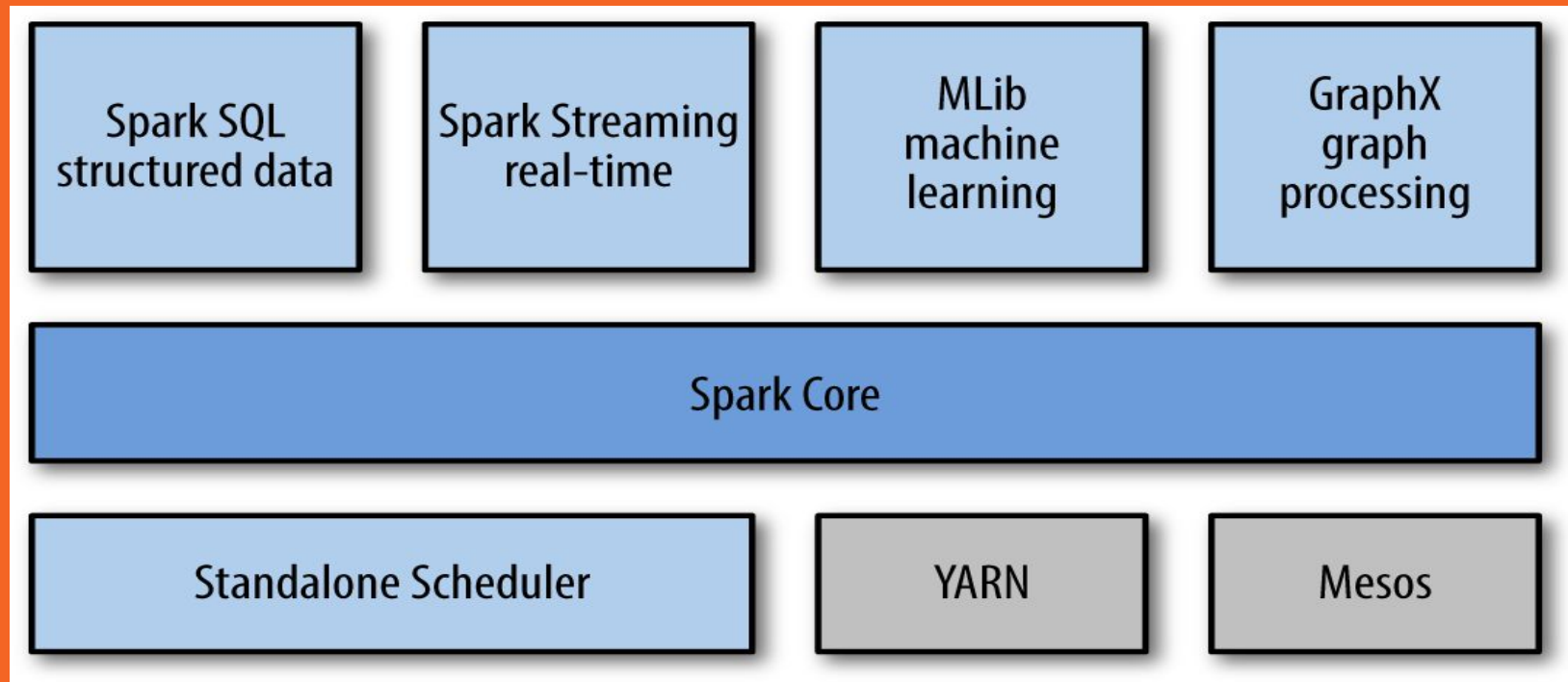
Completed Stages: 2

- ▶ Event Timeline
- ▼ DAG Visualization



# Spark Components

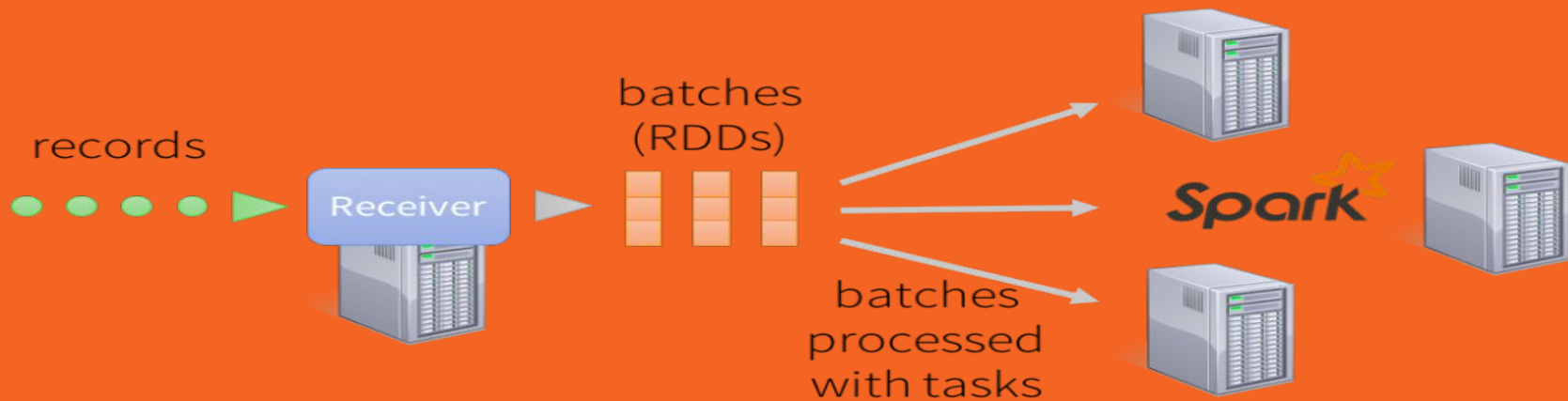
---



---

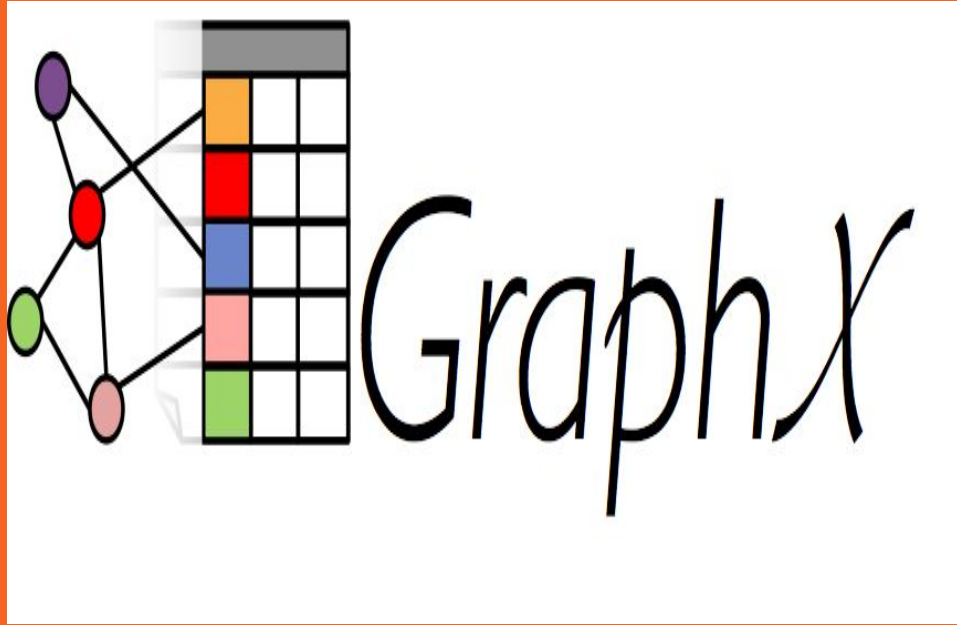
# Spark Streaming

*discretized stream processing*



records processed in batches with short tasks  
each batch is a RDD (partitioned dataset)

---



---

# Email spam and non-spam Classification demo

---

---

**Spark setup demo:**

**[https://spark.apache.org/  
downloads.html](https://spark.apache.org/downloads.html)**

---



## Important Links

---

<https://spark.apache.org/docs/latest/quick-start.html> → Getting Started with spark

<https://www.udemy.com/machinelearning/> → Start Machine Learning

machine learning with python o'reilly pdf → I will share this book on a page later

<https://www.udemy.com/apache-spark-with-scala-hands-on-with-big-data/> → for spark

---

---

# Questions

---