



Big Data : Where do I Start?

- Pathik Paul
- Ashok Sharma

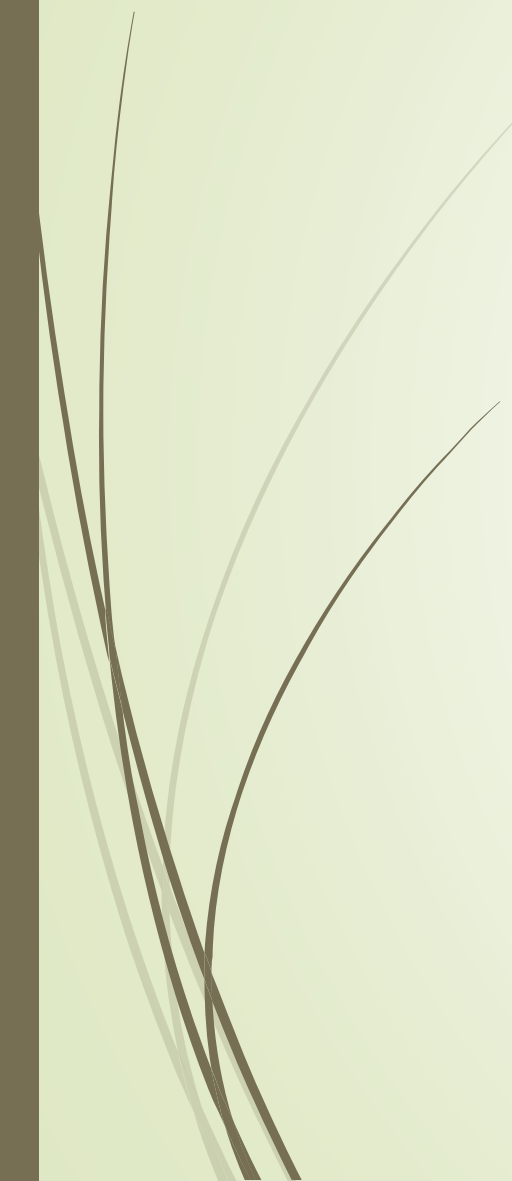


Suggested Timeline

- ▶ 7:00 - 7:15 - Introduction
- ▶ 7:15 - 7:30 - Recap & Can I move to big data
- ▶ 7:30 - 8:00 - Demo - Install & Run a Linux Machine (on Windows)
- ▶ 8:00 - 8:30 - Demo - Setup and Start Hadoop (Cloudera)
- ▶ 8:30 - 9:00 - Demo - Hadoop Brief introduction (Cloudera Exercises)



Why Move to Big DATA

- ▶ Why Not
 - ▶ Market is growing exponentially
 - ▶ Major changes happening in the industry
- 



Can I move to Big Data – Yes

- ▶ Shortage of resources
- ▶ Easy to Learn
 - ▶ Lots of free or low cost training options available
 - ▶ Attend Meetups
- ▶ Developer
 - ▶ If you know Java or Python or ... **SQL**
- ▶ Open Source
- ▶ Learn at home (Big Lab **Not** Needed)
- ▶ Get Certified
- ▶ Look for Internal Transfers

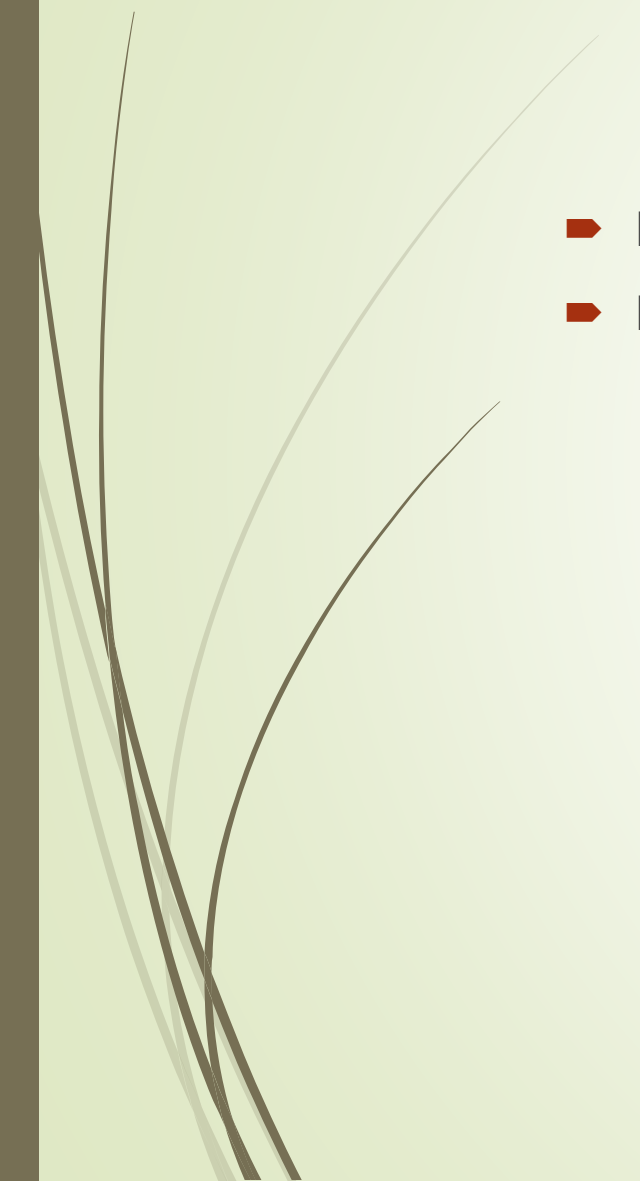


Roles and Jobs

- ▶ Administrator
 - ▶ Learn the basics
 - ▶ Take a Hadoop Admin course
 - ▶ Read the books
 - ▶ Get a Certification from Cloudera
 - ▶ Attend Meetups
- ▶ Developer
 - ▶ Focus on a Language + tool like Java + Map Reduce
 - ▶ Support it with Greenplum, Impala etc if you want the DB
 - ▶ Get a Certification from Cloudera
- ▶ Data Scientist
 - ▶ There are courses in the internet
- ▶ **Many More Options / Combinations available**



Recap: What is Big Data

- ▶ Data which cannot be processed by traditional programming means
 - ▶ Data which has some of the following characteristics / challenges
 - ▶ Volume
 - ▶ Velocity
 - ▶ Variety
- 



Recap: Databases

- ▶ From Flat File to RDBMS
 - ▶ NoSQL = Not Only SQL
 - ▶ Key Value
 - ▶ Column Oriented
 - ▶ Document Oriented
 - ▶ Graph Db
 - ▶ Massively Parallel Processing Databases
 - ▶ Greenplum
 - ▶ Many Others
- ▶ One size does not fit all
- ▶ Use the database that meets your needs




Big Data Sample Problem(s)

- ▶ Sears previously kept data anywhere from 90 days to two years
 - ▶ With Hadoop/Big Data they can keep everything online
- ▶ Google
 - ▶ How to store the web
 - ▶ How to update it on a regular basis
 - ▶ How to handle multiple types of data



Recap: Solution using Big Data

- Storage
 - GFS or HDFS
 - Replication
 - Knows where your data is stored
- Processing
 - Map
 - No Need to move the data to a central place
 - Perform processing Locally
 - Reduce
 - Summarize and present
- ****Low Cost Hardware**
- ****Automatic Fault detection and correction**



Recap: Hadoop

- ▶ Open Source
- ▶ Based on Google's White Paper


- ▶ Many Distributions available
 - ▶ Cloudera
 - ▶ Hortonworks
 - ▶ MapR
 - ▶ Many more



References



- <http://www.cio.com/article/2385690/big-data/5-reasons-to-move-to-big-data--and-1-reason-why-it-won-t-be-easy-.html>
- <http://www.datasciencecentral.com/profiles/blogs/10-reasons-why-big-data-analytics-is-the-best-career-move>
- <http://knowledge.sunstone.in/how-to-move-to-big-data-best-career-move/>
- <https://www.quora.com/Can-I-move-from-testing-to-big-data-Hadoop-development>
- <http://www.informationweek.com/it-leadership/why-sears-is-going-all-in-on-hadoop/d/d-id/1107038?>



Demo - Install & Run a Linux Machine (on Windows)

➤ <http://segintechnologies.com/setup-centos-virtual-machine/>



Demo - Setup and Start Hadoop (Cloudera)

- ▶ You will need to Sign Up to download the Machine
- ▶ http://www.cloudera.com/content/www/en-us/downloads/quickstart_vms/5-5.html
 - ▶ We used CDH5
 - ▶ Virtual Box
 - ▶ Start the Machine



Demo - Hadoop Brief introduction (Cloudera Exercises)

- ▶ Follow the tutorial
- 