ACADGILD

Presents

Introduction to Big Data and Hadoop2.x

Brief Intro About AcadGild:
CEO – Vinod Dham, Father of Pentium

- ACAD**GILD** is a technology education start-up which provides online courses in latest technologies like FrontEnd, FullStack, Big-Data, Android etc.
- Started by IIT/IIM alumni
- **Our aim** is to provide job ready skills to millions of high school and college graduates, and working professionals.

Course Objectives

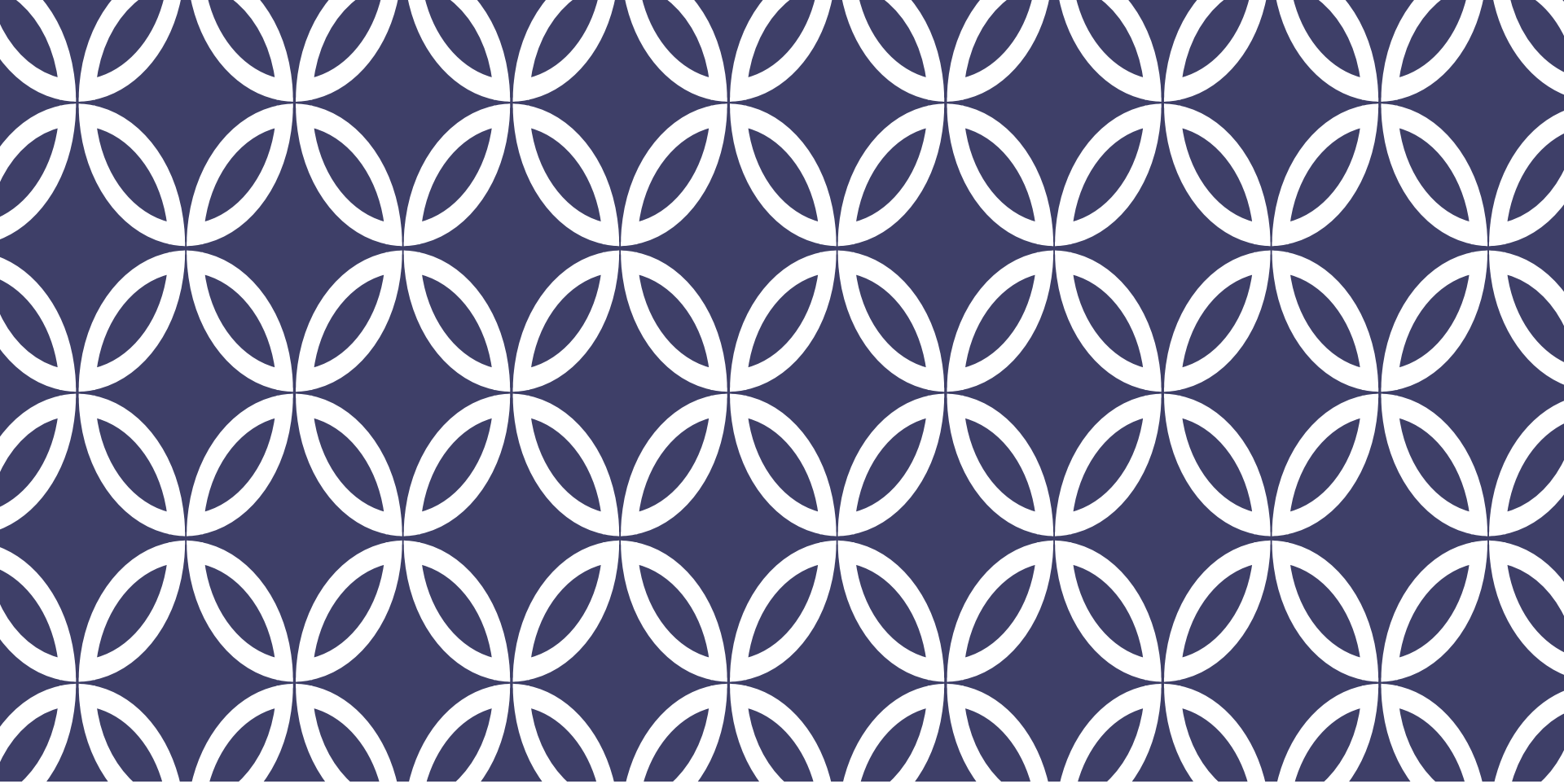
- Define & Describe Big Data with examples.
- Understand the Solution of Big Data – Hadoop Technology.
- Learn Hadoop Architecture.
- Creating our own single node and multi node cluster set up on linux platform.
- Learn Key Components of Hadoop Ecosystem in depth.
  - MapReduce, Yarn, HDFS, Pig, Hive, Hbase, and Oozie
- Data Loading Techniques of Flume & Sqoop.
- Assignments / Hands-on lab practice imparted with step-by-step guidance on tasks – to help you become an expert in writing your own programs on Hadoop2.x.

Towards the course end, you will apply your learnings by working on **two end-to-end real life projects on Hadoop!**

Sessions Break-Up

Session #	Module
1	How to Solve Big Data Problem
2	Hadoop Framework Description
3	HDFS
4	Hadoop Installation & EcoSystem
5 & 6	MapReduce
7	Pig
8	Pig Contd
9	Advanced Pig
10 & 11	Project1 – Enterprise Data Dictionary (EDD)
12	Advanced MR - Demo
13	Advanced MR – Demo Contd

Session #	Module
14	Introduction to Hive
15	Advanced Hive opertions
16	Hive UDF, UDF Demo using Hive, Thrift Server Demo, and Assignment
17	Introduction to HBase
18	HBase Architecture
19	Hands-On HBase CRUD Operations, Hbase Thrift Server, & Assignment
20	Apache Sqoop & Hands-On
21	Apache Flume & Hands-On
22	Introduction to Apache Oozie
23 & 24	Project 2: Text Mining over Hadoop to do Sentiment Analysis



Session 1 – How to Solve Big Data

Agenda – How to Solve Big Data

- Introduction to Big Data
  - What is Big Data
  - 3Vs of Big Data
- Sources of Data flood
- Exploding Data Problem
- Solution for Data exploding – Hadoop
- Evolutionary features of Hadoop
- Hadoop Timeline
- Who is using Big Data
- Job Trends in Big Data

What is Big Data?

3 Vs of Big Data

Volume
Data Size

- TeraBytes
- Records
- Transactions
- Table/Files

Data Complexity

Variety
Data Sources

- Structured
- Unstructured
- Semi-structured
- All of the above

Velocity
Speed of Change

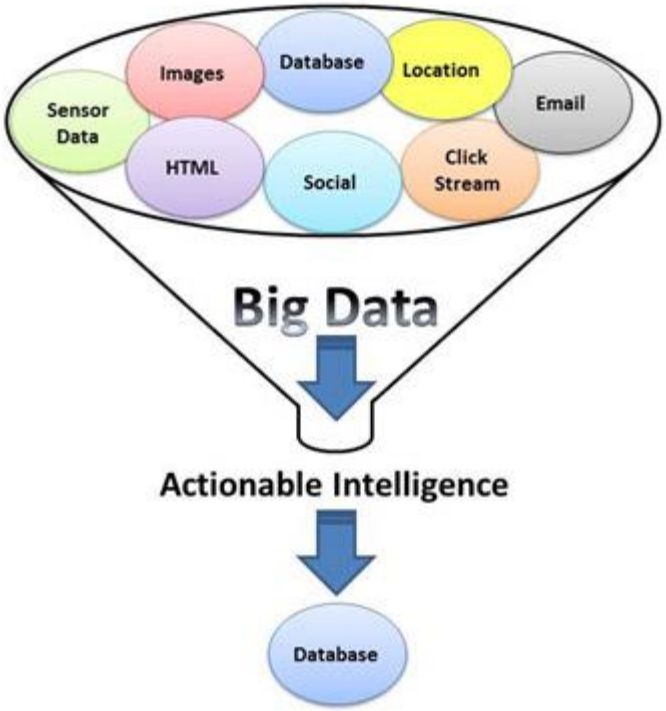
- Batch
- Near-Time
- Real-Time
- Streams

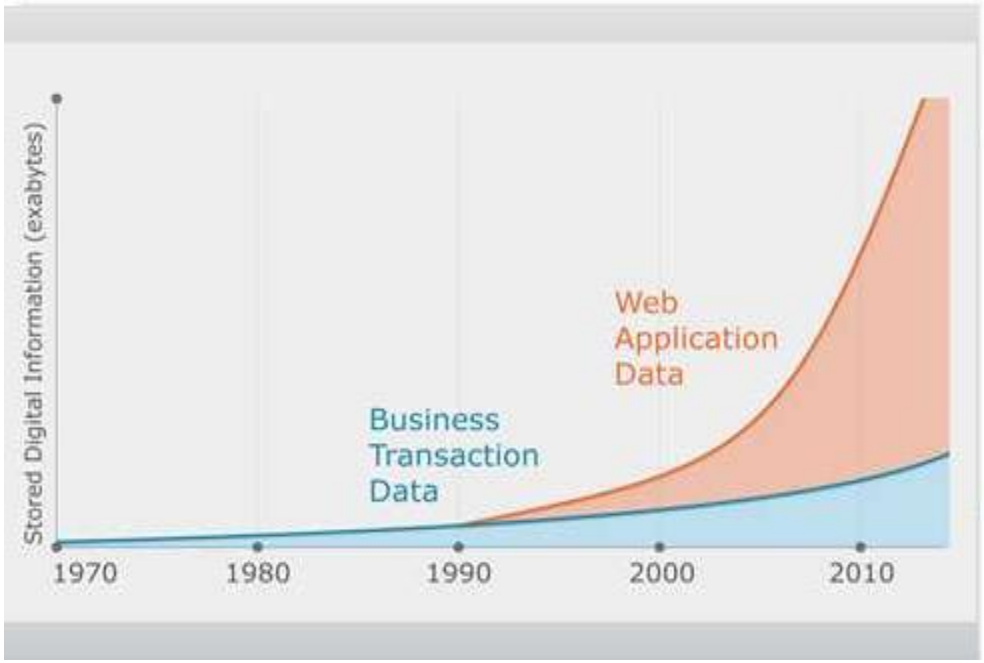
Sources of Data Flood

- This flood of data is coming from many sources.
- The new York stock exchange generates about 4-5 terabytes of data everyday.
- Facebook hosts more than 240 billion photos,growing at 7 petabytes of data everyday.
- Ancestory.com,the genealogy site stores around 10 petabytes of the data.
- The Internet Archive stores around 18.5 petabytes of data.
- The large Hadron collider near Geneva produces about 30 PetaBytes of data every year.

Exploding Data Problem

- Big Data constitutes - Large data sets in PBs & ZBs which cannot be processed by a single machine within expected time frame.





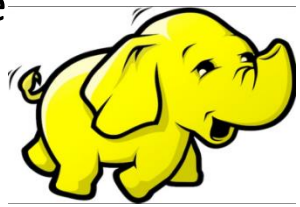
From the Pen of Eric Schmidt- Ex CEO GOOGLE

Every two days now we create as much information as we did from the dawn of civilization up until 2003, according to Schmidt. That's something like five exabytes of data.now

Solution for Data exploding - Hadoop

Need a new System:

- With new database management other than Relational Databases capable of handling unstructured as well as structured data.
- To process huge datasets on large clusters of computers than on a single system.
- To manage clusters in which:
  - Nodes fail frequently
  - Number of nodes keep changing
- Common infrastructure which is:
  - Efficient
  - Easy to use
  - Reliable

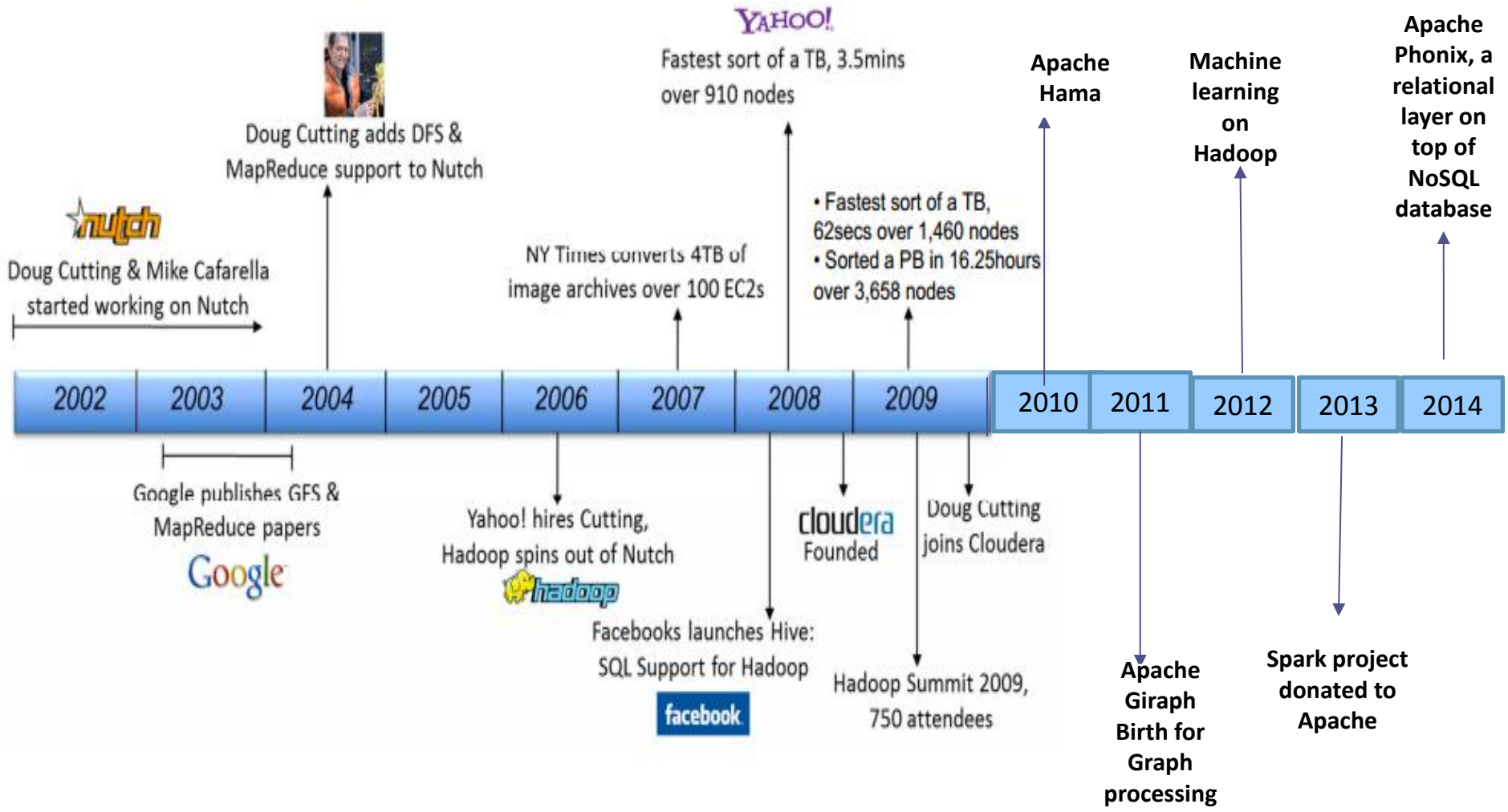


Hadoop is that new system !!

Over the years new processing patterns emerged:

Interactive SQL:	Impala has been integrated with Hadoop and Hive to build a new distributed query engine Tez to achieve low latency responses on SQL queries on hadoop.
Stream processing:	Streaming systems are Storm and Spark. Streaming has made it possible to run real time distributed computations on unbounded streams of data and send results to hadoop storage systems.
Search:	The Solr search engine can run on a hadoop cluster and can serve search queries from Indexes stored in HDFS.

Hadoop Timeline

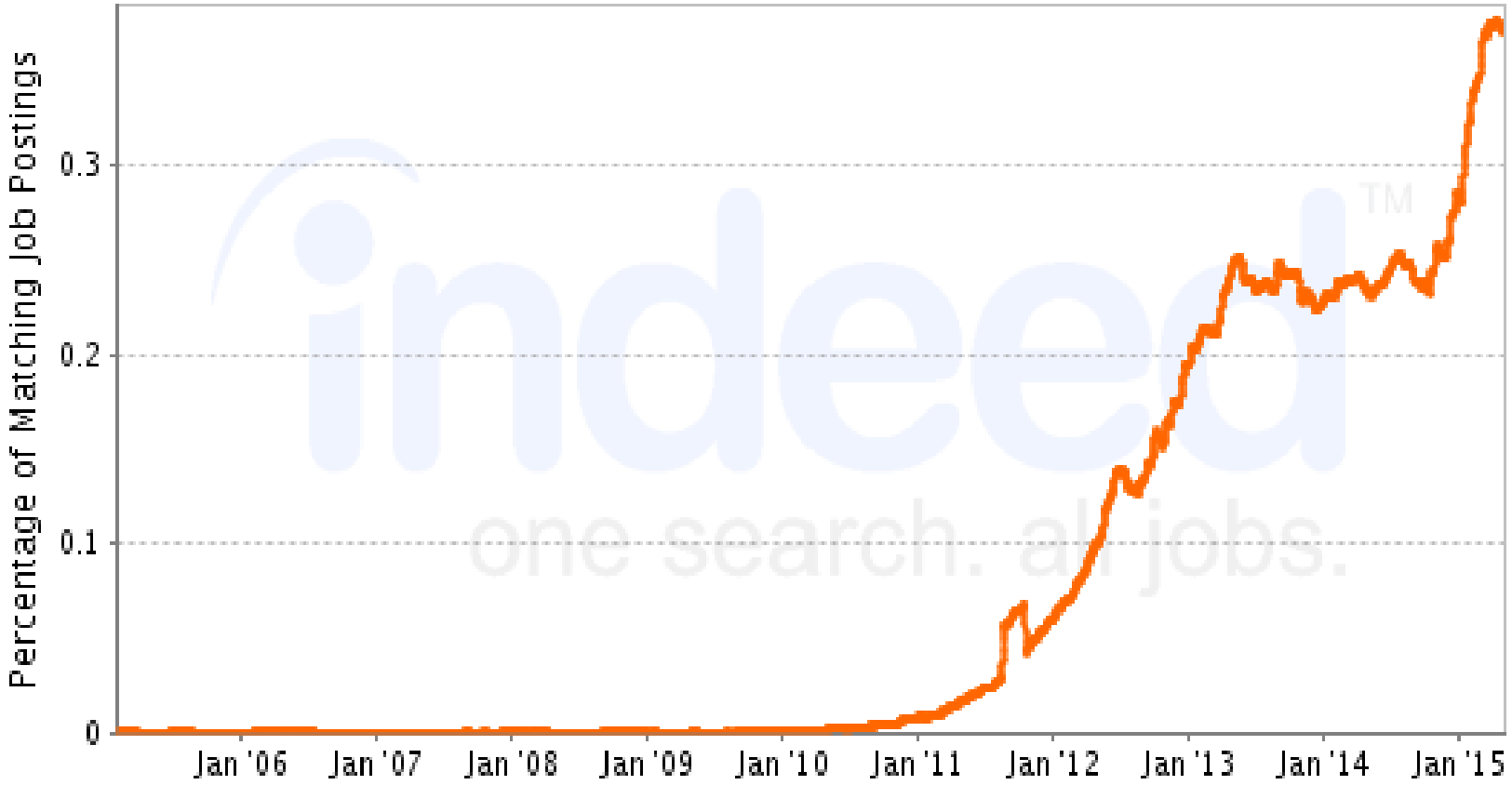


Case Study 1: Birth of Hadoop

Who is using Hadoop?

Case Study 2: Yahoo, Facebook, Apache HBase...

Job Trends in Big Data



Source: http://www.indeed.com/jobtrends/Big-data.html

Questions - Big Data

- What is Big data?
- How much is the Quantum of Big data, is it a constant or a variable?
- Why organizations are interested in Big data problems?
- Why data is increasing exponentially in the world wide web?
- Why scaling up is not a solution to Big data problem?

Topics in next Session(2)

Hadoop Framework Description:

1. Solving Big Data Problem
2. Hadoop Cluster
   - Introduction
   - Concepts
   - Why Hadoop Cluster?
   - Why Hadoop2 Came About?
   - How Hadoop Works
3. Hadoop 1.x Architcture
4. Progression from Hadoop 1.x to Hadoop-2.x
   - Core Components of Hadoop
   - NameNode Backup in Hadoop1.x
   - HDFS – High Availability Feature in Hadoop 2.x
5. Introduction to a YARN Application
6. Anatomy of a YARN Application
   - Phase I
   - Phase II

Get In touch with AcadGild

Contact Info:

- **Website : http://www.acadgild.com**
- **LinkedIn : https://www.linkedin.com/company/acadgild**
- **Facebook : https://www.facebook.com/acadgild**
- **Support: support@acadgild.com**