BIG Data Analytics
AWS Training
About Instructor

• Name: Kesav
• Total IT work experience: 20+ Years
• BIG Data Solutions Architect: 5+ Years
• DW & BI Solution Architect: 15+ Years
• Big Data Implementations Experience:
  – Hortonworks Hadoop echo system at Verizon
  – Amazon EMR echo system at Wallenius & Wilhelmsen Logistics
  – Cloudera Hadoop echo system at Bank of America Merrill Lynch
Students Introduction

• Please Introduce yourselves.
• Please let us know your reasons for taking this training, so that we can guide you better and help reach your goals.
Prerequisite

- Amazon AWS Essentials and basic knowledge
- Any database knowledge and basic SQL knowledge.
- Basic Linux/Unix commands & shell script knowledge.
Target Jobs

- BIG Data Analytics Developers
- Datalake ETL Developers
- Data Science jobs
Course Methodology

• Subject matter is delivered through:
  – Lecture and slide presentations
  – Software demonstrations
  – Class discussions
  – Hands-on labs
Quiz and tests

- At the end of each topic there will be Q&A session and Quiz.
- At the end of the course there will be a test evaluation.
Training Schedule

• Option 1 - Online Sessions :
  – 2 Weeks (excluding week ends) - Planned Schedule
  – Mon, Tue, Wed, Thu, Fri (8 pm – 10 pm EST)
  – Week ends (Students should continue Labs Practice)

• Option 2 – Onsite/Online Sessions :
  – 2 Weeks ends (excluding week days) - Planned Schedule
  – Saturday & Sunday (10 AM – 12 Noon & 2 - 5 pm EST)
  – Week days (Students should continue Labs Practice)
Course Content

• Module 1: Overview of AWS Essentials
  – Lab1: Introduction to AWS Lambda
• Module 2: Overview of Big Data & Hadoop Essentials
  – Lab2: Introduction to Amazon DynamoDB
• Module 3: Big Data Storage Solutions & Introduction to Amazon EMR
  – Lab3: Introduction to Amazon EMR
• Module 4: Ingestion, Compression, Storing and Processing Data
  – Lab4: Exploring Google Ngrams with Hive on Amazon EMR
• Module 5: Amazon Kinesis Streams & Spark Streaming
  – Lab5: Serverless Architectures with Amazon DynamoDB and Amazon Kinesis Streams with AWS Lambda
• Module 6: Amazon Kinesis Firehose – Leveraging for IOT Analytics
  – Lab6.1: Introduction to Amazon Kinesis Firehose
  – Lab6.2: Introduction to AWS Internet-of-Things (IoT)
• Module 7: Introduction to Amazon Redshift
  – Lab7: Introduction to Amazon Redshift
• Module 8: Introduction to Machine Learning
  – Lab8: Introduction to Amazon Machine Learning
• Module 9: Visualizing and Orchestrating Big Data
  – Lab9: Using Tableau Desktop with Amazon Redshift
• Module 10: BIG Data Analytics - Business Scenarios & Case Studies
  – Lab 10: Case Study – Advanced Amazon Redshift: Analytics and Amazon Machine Learning
What is Data warehouse

• Data warehouse is the main repository of organization's historical data (corporate memory) for management's decision support systems.

• Data warehouse is optimized for reporting and analysis (OLAP) systems while operational systems (OLTP) are optimized for simplicity and speed of modification through database normalization and an entity-relationship model.

• Data warehouses is deormalized, summarized and/or stored in a dimension-based model. The critical factor leading to the use of a data warehouse is that a data analyst can perform complex queries and analysis on the information without slowing down the OLTP systems.
What is Business Intelligence

• Business intelligence (BI), is an umbrella term that refers to a variety of software applications used to analyze an organization’s raw data.

• BI as a discipline is made up of several related activities, including data mining, online analytical processing, querying and reporting.

• Companies use BI to improve decision making, cut costs and identify new business opportunities.

• BI is more than just corporate reporting and more than a set of tools to coax data out of enterprise systems.

• Although BI holds great promise, executives have to ensure that the data feeding BI applications is clean and consistent so that users can trust it.
Business Intelligence

• Examples:
  – Show me the most profitable customers
  – Show me the most effective promotions
  – Show me the sales for each area by month
  – Show me the products that are not profitable
  – Show me the customers most likely to switch
  – Compare the sales/pipeline for this quarter vs. last quarter
DW/BI Benefits

- Access mission critical information
- Improved business efficiency and productivity analysis
- Visibility into your entire business (Internal and External)
- Improved visibility about Customers, Products, etc
- Make Informed decisions and respond to changes more easily
What is BIG Data

• Today the term big data draws a lot of attention, but behind the hype there's a simple story.
• For decades, companies have been making business decisions based on transactional data stored in relational databases.
• Beyond that critical data, however, is a potential treasure trove of non-traditional, less structured data: weblogs, social media, email, sensors, and photographs that can be mined for useful information.
• Decreases in the cost of both storage and compute power have made it feasible to collect this data - which would have been thrown away only a few years ago.
BIG Data Definition

• The most common explanation of Big Data is defined by the three “V’s” – Velocity, Volume, and Variety.
• Velocity is the speed at which data is being generated via multiple channels like social media and networks.
• The other dimensions are the advent of unstructured and semi-structured data that is being created like Twitter feeds, Facebook entries, LinkedIn updates, audio files, video files, various types of documents and rich transactional data.
• The combination of the three V’s and new data formats are the factors that uniquely define Big Data.
BIG Data Architecture

Data
- Master & Reference
- Transactions
- Machine Generated
- Social Media
- Text, Image, Video, Audio

Acquire
- DBMS (OLTP)
- Files
- NoSQL
- HDFS

Organize
- ETL/ELT
- ChangeDC
- Real-Time
- Message-Based
- Hadoop (MapReduce)

Analyze
- ODS
- Warehouse
- Streaming (CEP Engine)
- In-Database Analytics

Decide
- Reporting & Dashboards
- Alerting & Recommendations
- EPM, BI, Social Applications
- Text Analytics and Search
- Advanced Analytics
- Interactive Discovery

Management, Security, Governance

Specialized Hardware
- Big Data Cluster
- High Speed Network
- RDBMS Cluster
- In Memory Analytics
BIG Data Analytics Benefits

• A broad range of applications for BIG Data Analytics helping companies rack up impressive ROI figures.

• More and more companies are looking to include non-traditional yet potentially very valuable BIG data with their traditional enterprise data in their business intelligence analysis to identify cost-cutting ideas, uncover business opportunities, react quickly to retail demand and optimize prices.

• To derive real business value from big data, need the right tools to capture and organize them alongside existing data to find new insights and capitalize on hidden relationships.