Pavan(Data Engineer)

| 636 565 6150| pavanreddyfordata@gmail.com

# Summary

A Data Engineer with newly acquired skills, an insatiable intellectual curiosity, and the ability to mine hidden gems located within large sets of structured, semi-structured and unstructured data. Able to leverage a heavy dose of mathematics and applied statistics with visualization and a healthy sense of exploration. The skills which I got might be helpful for you and please look at my profile for at least 10 seconds.

## professional summary

* Around 6+ years of experience in Analysis, Architecture, Design, Development, Testing, Maintenance and User training of software application which includes over 5+ Years in BigData (Spark, Hadoop and HDFS) environment and around 1 Year of experience in **JAVA/J2EE.**
* **5 years (almost)** hands on experience on Hadoop (**HDFS, Map Reduce, PIG, HIVE, and SQOOP** etc.).
* **2 years (almost)** hands on experience on **Spark, Scala, Kafka, AWS, Hbase**.
* Developed end-to-end data pipelines which include **data extraction**, **data ingestion**, publishing data on **Tableau** server, and automation of pipelines under fast paced Agile Scrum environment
* Optimized pipelines in Data lake by implementing Partitioning, and bucketing concepts for improving performance. Accomplished and Improved performance of Spark Pipelines by implementing repartition to manage resources efficiently
* **HDPCD Certified Spark Developer** (Verification link- <http://bcert.me/sdlhgtfy>)
* Experience in Testing solutions with the **Microsoft Azure including HDInsight**.
* Migration from different databases (i.e. **Oracle**, DB2 and **MYSQL**) to **Hadoop** and **Spark** with **NoSQL** databases.
* **NoSQL** database experience with **HBase**and good exposure of **Cassandra**.
* Proficient in using **Cloudera Manager**, an end to end tool to manage Hadoop operations
* Experience using **Sqoop**to import data into **HDFS** from **RDBMS** and vice-versa.
* Expertise in **Unix**-based operating systems
* Expertise in developing **Spark** programs that includes **data processing**, **data management** etc.
* Experience creating real-time data streaming solutions using **Apache Spark core, Spark SQL, Kafka and spark streaming.**
* Hands on Experience in **Data Modelling (Erwin, Visual studio),Data Analysis**, **Data cleansing** and Entity Relationship diagrams (ERD).
* Experience in **metadata** maintenance and enhancing existing **Logical** and **Physicaldata models**.
* Extensive experience with **ETL** and Query with big data tool like Hive QL.
* New data science and **Machine learning** skills to derive actionable insights in industry and beyond.
* Have good interpersonal, communicational skills, strong problem-solving skills, Strong analytical and judgment techniques.

# Experience

## Client: Peridot solutions  july 2016 to present

## Role: Data Engineer

**Project:**The Project have both development and production environment, my duties focused on maintaining, and optimizing the existing databases; designing and implementing new database for various projects according to the requirement.

**Technology stack used -** Spark/Scala, Spark streaming, Kafka, Hbase, AWS, HDFS.

**Responsibilities:**

* Designed and deployed a Spark cluster and different Big Data analytic tools including **Spark, Kafka streaming, AWS andHBase with** Cloudera Distribution.
* Configured deployed and maintained multi-node Dev and Test **Kafka**
* Integrated **kafka**with **Streaming ETL** anddone some required ETL on it to extract the meaningful insights.
* Developed application components interacting with **Hbase**.
* Performed optimizations on **Spark/Scala**.
* Used the **Kafka** producer app to publish clickstream events into the **Kafka topic and** laterexplored the data with**sparkSQL**
* Processed raw data at scale including **writing scripts**, **web scraping**, **calling APIs**, **write SQL queries**, etc
* Importing streaming logs and aggregating the data to **HDFS and MYSQL** through **Kafka**.
* Exploring with the Spark improving the performance and optimization of the existing algorithms in Hadoop using **Spark Context, Pyspark, Spark-SQL, Data Frame, Pair RDD's** and **Spark YARN**.
* Implemented **Machine learning** algorithms to optimize electrode targeting and parameter settings for deep brain stimulation.
* Developed custom Machine Learning (ML) algorithms in Scala and then made available for **MLIB** in **Python** via **wrappers**
* Developed Spark code using **Scala** and **Spark-SQL/**Streaming for faster testing and processing of data.
* Imported data from different sources like **HDFS, MYSQL** andother sources through **Sqoop** and **kafka** to import streaming logs into **Spark RDD**
* Performed visualization using SQL integrated with **Zeppelin** on different input data and created rich dashboards
* Performed transformations, cleaning and filtering on imported data using **Spark-SQL** and loaded final data into HDFS and MYSQL database.
* Involved in production support and enhancement development.

**Environment:** Hadoop, Spark, Pyspark, Spark-SQL, HDFS, MapReduce, Hive, Sqoop, Kafka, HBase, Oozie, Spark - Streaming/SQL, java, SQL Scripting, Linux Shell Scripting, Zeppelin.

## Client: SEARSseptember 2015 to July 2016

## Role: Hadoop developer

**Project:**The project objective is to extract all the sales and the products sold across all the Sears Retail stores and load it into the data warehouse using the Big Data components. It also involves loading of data from traditional databases to HDFS using various Big Data technologies

**Responsibilities:**

* Developed different **MapReduce** applications on **Hadoop**.
* Mining the location of users on social media sites in semi supervised environment on Hadoop cluster using Map Reduce.
* Implementing single source shortest path on Hadoop cluster.
* Involved in loading and transforming large sets of Structured, Semi-Structured and Unstructured data and analysed them by running **Hive** queries and **Pig scripts.**
* Evaluated suitability of Hadoop and its ecosystem to the above project and implemented various proof of concept (POC) applications to eventually adopt them to benefit from the Big Data Hadoop initiative.
* Estimated Software & Hardware requirements for the Name Node and Data Node & planning the cluster.
* Participated in requirement gathering from the Experts and Business Partners and converting the requirements into technical specifications.
* Extracted the needed data from the server into **HDFS** and Bulk Loaded the cleaned data into **HBase**.
* Written the Map Reduce programs, **Hive UDFs** in Java where the functionality is too complex.
* Involved in running Hadoop jobs for processing millions of records of text data.
* Involved in loading data from **LINUX** file system to HDFS.
* Prepared design documents and functional documents.
* Based on the requirements, addition of extra nodes to the cluster to make it scalable.
* Developed HIVE queries for the analysis, to categorize different items.
* Assisted application teams in installing Hadoop updates, operating system, patches and version upgrades when required.
* Designing and creating Hive external tables using shared meta-store instead of derby with partitioning, dynamic partitioning and buckets.
* Given POC of **FLUME** to handle the real time log processing for attribution reports.
* Maintained System integrity of all sub-components (primarily HDFS, MR, HBase, and Hive)..

**Environment**:Hadoop, HDFS, MapReduce, Yarn, Hive, PIG, Oozie, Sqoop, HBase, Flume, Linux, Shell scripting, Java, Eclipse, SQL.

## Client:bank of america december 2011 to august 2014

## Role: Hadoop developer

**Project:**I was part of the DST(Mainframe) decommission project involved in moving the data from the mainframes to HDFS.

**Responsibilities:**

* Requirement discussions, design the solution.
* Estimated the Hadoop cluster requirements
* Responsible for choosing the Hadoop components (**hive, pig, map-reduce, Sqoop, flume** etc)
* Responsible for building scalable distributed data solutions using Hadoop.
* Hadoop cluster building and ingestion of data using Sqoop
* Imported streaming logs to HDFS through **Flume**
* Used Flume to collect, aggregate, and store the web log data from different sources like web servers, mobile and network devices and pushed to HDFS
* Developed Use cases and Technical prototyping for implementing Hive,and Pig.
* Worked in analyzing data using Hive, Pig and custom MapReduce programs in Java.
* Implemented partitioning, dynamic partitions and buckets in HIVE
* Installed and configured Hive, Sqoop, Flume, Oozie on the Hadoop cluster.
* Involved in scheduling Oozie workflow engine to run multiple Hive and Pig jobs.
* Tuned the Hadoop Clusters and Monitored for the memory management and for the Map Reduce jobs.
* Responsible for Cluster maintenance, Adding and removing cluster nodes, Cluster Monitoring and Troubleshooting.
* Developed a custom Framework capable of solving small files problem in Hadoop.
* Deployed and administered 70 node Hadoop clusters. Administered two smaller clusters.

**Environment:** Map Reduce, HBase, HDFS, Hive, Pig, Java, SQL, Cloudera Manager, Sqoop, Flume, Oozie, Java (JDK 1.6), Eclipse.

## Client: hsbcBank jan 2011 to dec 2011

## Role: java developer

**Project:** The main application is a secure web-based application allows the users to Register and provide information about Retirement plan like 401(k), IRA, Other Defined Benefits. This application mainly used to track participant and retiree data for the customer, including plan information such as status, service and earnings, and participant contact data. It also used to handle all reporting details, make benefit payments to retirees and take care of all Federal, State and local withholding requirements.

**Responsibilities:**

* Involved in various stages of Enhancements in the Application by doing the required analysis, development, and testing.
* Prepared the High and Low-level design document and Generating Digital Signature
* For the registration and validation of the enrolling customer developed logic and code.
* Developed web-based user interfaces using **J2EE** Technologies.
* Handled Client-Side Validations used **JavaScript** and
* Used Validation Framework for Server-side Validations
* Created test cases for the Unit and Integration testing.
* Front-end was integrated with Oracle database using JDBC API through **JDBC-ODBC** Bridge driver at server side.

**Environment:** Java Servlets, JSP, JavaScript, XML, HTML, UML, Apache Tomcat, Eclipse, JDBC, Oracle 10g.

# Skills & Abilities

## Technical Skills

**Hadoop/Big Data Technologies: Spark-Scala, Kafka, Spark Streaming, Mlib, Sqoop, Hbase, HDFS, Map Reduce, Pig, Hive, Zeppelin**

**(**Distributions: **Data Bricks**, **Horton works** and **Cloudera)**:

Programming Languages and Scripting : Java (JDK 5/JDK 6), C/C++, Python, **Scala**, HTML, SQL

Operating Systems : UNIX, Windows, LINUX, Mac OS X

Application Servers : IBM Web sphere, Tomcat

Web technologies : JSP, Servlets, JDBC, Java Script, CSS,

Databases : Oracle9g/10g & MySQL 4.x/5.x, Hbaseon AWS-s3 and HDFS

Data Modelling : Erwin, Visual Studio

Development Methodologies : Agile Methodology -SCRUM, Hybrid.

# Education

## Master of science | california state university fullerton | USA

Major: Computer Science

## Bachelor OF technology | JNTUH | India

Major: Computer Science