SHALU MISHRA

Email: shaluit3076@gmail.com
GitHub: https://github.com/shaluMishri
Mobile: +91 7755922094 / 8765179315
Location- Pune

Professional Summary

- Vibrant and Self Motivated with more than 2 years 10 month of experience in Data Analytics, Big Data Ecosystem related technologies and Cloud Services with software development, implementation and deployment in **Infosys Limited** Pune.
- Experience in Hadoop Eco-system (Hortonwork and Cloudera Enterprise Distribution), as well as Cloud Frameworks (Amazon Web Services and Azure).
- Worked on Big Data technologies and tools like Hadoop, Map Reduce , HDFS, Hive, Pig, Apache Spark, Oozie, Kafka, Solr , Hbase and Mongo DB, distributed computing.
- Hands on experience with Oozie for job workflow scheduling.
- Experience in analyzing the data using Spark SQL, HiveQL, Pig Latin, Hbase, Hive UDF.
- Experience in importing & exporting data using Sqoop from HDFS to RDBMS and vice-versa.
- Good Exposure in Java, Web Services,
- Knowledge in Informatica, Cognos, Tableau, SSRS and SQL Server in training.
- Strong track record of understanding and interest in current and emerging technologies demonstrated through training, job experience and / or industry activities.
- Ensuring the high quality work products under pressure within deadlines and adhering to the functional and technical designs and standards.
- Always ready to explore, learn and understand newer business domains and technology.

Skillset

Cloud Services	Amazon Web Services (AWS S3, AWS EC2, AWS EMR, lambda, IAM,VPC) , AZURE
Big Data Technologies	HDFS, Map Reduce, Hive, Hbase, Yarn , Oozie, Pig, Kafka, Solr, and Zookeeper, Apache Spark
Programming Languages	Python, Java, JavaScript, Shell Script HTML
Tools and IDE	Eclipse, Maven(Nexus Repository Setup),GitHub, Jenkins
BI Reporting Tool	Jmeter, Cognos

Trainings & Certification

- Undergone training on Big Data & Business Intelligence from Infosys Ltd, Mysore.
- Microsoft Certified 70-533 Implementing **Microsoft Azure Infrastructure Solutions.**
- Certification ADM 200- Cluster Administration Essentials: Install a MapR Cluster.
- Certification ADM 201- Hadoop Operations: Cluster Administration.
- Certification ADM 202- Cluster Administration Essentials: Data Access and Protection.
- Certification ADM 203- Cluster Administration Essentials: Cluster Maintenance.
- Certified in Hadoop, Mongo DB, Hive from Infosys.

Academics

- Graduate (Bachelor of Technology) in Information Technology from UPTU in **2014**
- Intermediate, ISC in **2010**
- High School, ICSE Board in **2008**

Work Experience

Suntrust Banks, Inc
Role: Hadoop developer   **(Jul17 – Till Now)**

Description: The aim of this project is to analyze the data which is coming from multiple sources into the Hadoop data lake. Created programs to process large volumes of data through a lot of prepay concepts, which analyze, produce suspect claims and it helps to generate datasets for visualization. As a Hadoop developer, involved in maintaining the huge data and designing developing predictive data models for business users according to the requirement also to specify design for automating account authenticity prediction based on machine learning algorithm model.

Responsibilities:
- Hadoop Cluster setup in Amazon web services EMR with Apache Spark and Leveraged AWS S3 as storage layer for HDFS.
- Hive tables were created on HDFS to store the data processed by Apache Spark in Parquet format.
- Used Sqoop for importing and exporting data from MySql into HDFS and Hive.
- Imported Hive tables into Spark SQL context and converted into RDDs.
- Used Data Frames and Datasets APIs for performing analysis on Hive tables.
- Developed PySpark and Spark SQL scripts using python to perform transformations and actions on RDDs in spark for faster data Processing.
- Generate automated solution for populating prediction score for accounts authenticity checking on review comments / analytics comments using Pyspark ML and MLib.
- Automating account authenticity prediction based on SVM (Support Vector Machine) machine learning algorithm model.
- Developed Python code using PySpark ML,Mlib for prediction of analysis report of complaints logged and pyspark code for model on training Set.
- Automated and scheduling the Hadoop Jobs using oozie for time frequency and data availability.
- Monitoring of Hadoop Cluster using Cloudera Manager.
- Load and transform large sets of semi structured and unstructured data that includes sequence files and xml files and worked on Avro and Parquet file formats using compression techniques like Snappy, Gzip and Zlib.
- Used Amazon Simple Storage Service(S3) for storing and accessing data to hadoop cluster.

Environment: HDFS, Yarn, MapReduce, Sqoop, Oozie, Hive, HBase, Spark Machine Learning, Spark Streaming, Spark SQL, PySpark, Python, Eclipse, Cloudera, AWS, S3, EC2.

Underwriter laboratory
Role: Hadoop developer   **(Dec16 – Jun17)**
Client: Global independent safety Science Company with more than a century of expertise innovating safety solutions from the public adoption of electricity to new breakthroughs in sustainability, Renewable energy and Nanotechnology.

Description: Build a digital platform which provides Small Businesses with information and insights about their business, customers. It also offers tools to help businesses take action. Development includes data acquisition from various source systems to HDFS, transform, categorize and aggregate using business rules, create different SOLR services to fetch input data from Hbase to expose at user portal. Pulling the data from various data sources into Hadoop platform using Cloud Framework and standardized all data through a series of master data management processes to reach client goal.

Roles and Responsibility:
- Configured and Deployed Fully-Distributed And high availability Apache Hadoop Cluster on Amazon Web Service Cloud. Setting up the Cloud (Cluster Launcher VPC, Subnet, NAT, BASTION HOST, Master, Worker, and Internet Gateway) and high Availability Cloud Formation.
- Monitoring and managingl Hadoop services through AWS WebUI. Working Knowledge of Cloudera Manager.

- Generic Injection: Extracting excel and creating Custom Input json and table extraction using map reduce. Transformation of raw json to modelled json and loading modelled json to HBase and added to Solr.
- Application Automation for Hbase and Solr. Built SOLR collection and create indexers to expose aggregated data to customer portal. Using lily Indexer with hbase to load data in Solr.
- Loading data on AWS S3 cloud storage from multiple sources.
- Performance Testing for Solr/HBase backup and restore on cluster. Oozie scheduler with Cron Job Setting
- Generating Test reports using Jmeter.
- EMR cluster creation using AWS CLI with bootstrap script which includes installation of maven and Jenkins.
- Writing Hive Queries to transform data as per business requirement and push data into Elastic search to create dashboard using Kibana (4.4.0).
- Encryption Techniques on data moving to AWS S3 from various resources.

Tools and Technologies: Azure**,** Hortonwork Data Platform ,AWS-EC2, AWS-S3, RDS, EMR, Elastic Beanstalk, AWS Lambda, AWS IAM, Map Reduce, HDFS, HBase, Solr, Hue, Oozie, Java, Web Services, Putty, Git bash.

Client: American Xpress
Role : Hadoop developer   **(Mar16 – Dec16)**

Description: The use case of the project is tracks call logs, dropped call and calculates the peak hours and categorizes user based on the call volumes for offers.

Roles and Responsibility:
- Requirement gathering, Requirement Analysis and monitoring Hadoop jobs. Worked on Real time Streaming and transformation of streams and makes the results available system-wide.
- According to the call plan, offers are being given to user through SMS.
- Using Kafka to subscribe the topic from multiple web servers and consume the data from broker and stored the data into HDFS.
- Create Web services using Rest API to connect Kafka topics to read the call logs ,location, plans to monitor and produce messages ,generate output summaries.
- Using Java API for converting raw data to DStream and sending data to spark for processing.
- Stateful Analysis with spark streaming, to find out the peak hours for each call type and categorizing user on the basis of call plans.
- Extracted data are processed using smart data cleanse to apply business specific data clean-up and de-duplication using Spark utilities.
- Data are further send for report generation using Hive, Tableau.

Tools and Technologies: Java, Web Services Rest API, Shell Script, Map Reduce, HDFS, Kafka, Spark Streaming

Client: American Xpress
Role : Hadoop developer   **(Jul15 – Mar16)**

Description: The use case of the project is Merchant Acquisition Tracking Log Incentives and Compensation (MATLIC) which tracks salary and seller incentives and compensation based on the different business scenarios related to the opportunities for all the merchants which has on boarded into the Client User Interface.

Roles and Responsibility:
- Requirement gathering and analysis and translate complex functional and technical requirements into detailed design.
- Analyzed the source data, worked on data cleansing by means of removing null values, target table structures. Transformed raw data from several data sources into baseline data by developing Pig scripts and loaded the data into HBase tables.
- Design and implemented Pig (UDFs) for evaluation, filtering, and transforming large volumes of data with respect to business requirement.
- Developed HIVE scripts with complex Joins, Partitions for Transformed Data.

- Developed Shell Script to perform Operation on HDFS, and to launch Oozie scheduler.
- Data Orchestration by Oozie workflow to automate the tasks and monitor Hadoop jobs.
- Performances Testing and Load Testing while loading data.

Tools and Technologies: Java, Web Services, Shell Script, Map Reduce, Hive, Pig, Oozie, Hbase

POC: MASTER HEALTHCARE
Description: Master Health was designed and developed for American Cancer Hospitals (A Super specialist Hospitals offering distinguished medical and Theraphatic services around the clock). This project is basically providing sophisticated software for automation of the various transactions in the hospital .It includes several modules like Security, Registration, Doctor Appointments, Inpatient, Outpatient, Investigative lab Services, Operation Theatre, Finance and Accounting with MIS Reports.

Roles and Responsibility:
- Gathered the business requirements and documentation with analysis.
- Feasibility Analysis – Evaluating the feasibility of the requirements against complexity and time lines.
- Created Hadoop cluster with different tools like pig, Sqoop, oozie, hbase , hive .
- Analyzing structured and unstructured data. Loading the data to HDFS.
- Developed PIG Latin scripts to extract data from source system.
- Developed Hive queries to analyze data.

Tools and Technologies: Putty, Eclipse, maven, winscp, Hadoop, Hbase, HDFS, Hive, Pig, Sqoop.

Short-Term Projects:
- **Time Series Analysis Spark-Cassandra:** In Utilities Solr and Wind Turbines generate the time-series data and which can be used monitor the power quality and out. Cassandra is the one of the good choice to store and manage the time-series data and Spark is good processing Engine. We wanted to see how best we can leverage these technologies and implement a POC
- **Web Portal Creation:** We need to build web portal for the document generation based on Web UI entries. Web page should be rendered based on the configuration mentioned in XML. Generated document should be in PDF and Word format. Also, the approval workflow of document need to be created and managed.

Extra-Curricular Activities

- Represented **India** and won Medals in **International Gymnastics competitions**.
- **State Gold Medalist** in Gymnastics.
- Robotics third prize winner at College level event.
- Part time teaching to students at primary level without fees.

Personal Details

Name	**:** Shalu Mishra
DOB	**:** 20-Jan-1992
Marital Status	**:** Single
Address	**:** 415 Mumfordganj Allahabad, Uttar Pradesh 211002
Hobby	: Gymnastics, Reading, Playing Guitar

I hereby declare that information provided is authentic and true to the best of my knowledge.

Date: **Shalu Mishra**