**Professional Summary:**

8+ years of Cumulative IT experience in software development, ETL and Big Data. Excellent understanding and Hands on experience in Hadoop Framework.

**Experience Summary:**

* 3+ years of experience as Big Data developer in Hadoop Ecosystem & Different Big Data analytical tools.
* Writing UDF (User Define Functions), UDAF (User Define Aggregate Functions), UDTF (User Define Table Functions) for Hive and Spark.
* Writing RegexSerDe to process log data and store it into Hive warehouse.
* Hive optimization with Cost Based Optimization, Map side Joins, Strict Mode, Vectorization, ORC.
* Creating Internal(Managed) and External Tables in Hive to improve query optimization.
* Tuning Catalyst optimizer for spark to optimize query execution plan.
* Tuning Garbage Collector (Eden, Survivor 1, Survivor 2) for Memory optimization.
* Memory optimization with Kyro and Custom Serialization techniques.
* Providing High Availably with Zookeeper.
* Creating Kafka Producers and Consumers with Java Programs.
* Creating Kafka Topics and registering those topics with zookeeper.
* Providing Security with Kerberos authentication techniques.
* Spark Performance optimization with GC.
* Experience in developing applications using Hadoop ecosystem like Spark Streaming, Spark SQL, Hive, Flume, Kafka, Sqoop and HBase.
* Writing Python Scripts for Data Scrape from Sources and Store it on S3 Buckets and run analytics.
* Extensive knowledge of Hadoop & Spark architecture and core components.
* Having Knowledge in Containerizing technologies like Docker.
* Having Knowledge in Orchestration Platform Like Kubernieties.
* Experience in writing queries for moving data from HDFS to Hive and analyzing the data using Hiveql.
* Experience in importing and exporting data using Sqoop from Relational Database Systems (RDBMS), Hive, HBase to HDFS, Experience in creating hive context in spark-shell.
* Creating connections between Kafka and flume using connection pipelines.
* Created Sqoop jobs with incremental load to populate Hive External tables.
* Experience in creating Kafka connectors to ingest data into Spark Streaming
* Creating SNS (Simple Notification Services) in AWS.
* Analyzed large amounts of data sets writing Hive Queries.
* Knowledge of Hive Partitioning, Bucketing, Join optimizations and Query Optimizations
* Involved in converting Hive SQL queries into Spark SQL
* Knowledge of Spark core, Spark SQL, Spark Streaming, Data Frames, RDD’s, Spark Python API modules of Spark, Scala for Spark and Spark Data frames.
* Experience in using DStreams, Accumulators, Broadcast variables, RDD caching for Spark Streaming
* Knowledge of Spark SQL UDF, RDD Partitioning.
* Experience in Extraction, Transformation & Loading (ETL) of data with different file formats like CSV, Text files, Sequence files, Parquet, XML, JSON and Avro files based on business requirements.
* Familiarity with the Hadoop architecture, design of data ingestion pipeline, data mining and modeling, advanced data processing and machine learning.
* Ability to plan, manage, motivate and work efficiently in both independently or in a team effort.
* Good exposure in overall SDLC including requirement gathering, development, testing, debugging, deployment, documentation and production support.
* Experience in working with software methodologies like Agile, Prototype & Waterfall.

**TECHINICAL SKILLS**

|  |
| --- |
| Hadoop Ecosystem : Spark 2.0, Hive, MapReduce, Sqoop, Flume, Kafka. |
| Apache Spark : Spark SQL, Spark Streaming, Structured Streaming. |
| Languages : Scala 2.11,Python 2.7, SQL, Unix Shell Scripting. |
| Java technologies : Java, J2EE, Servlets, JDBC, Hibernate. |
| ETL : Informatica. |
| Databases : MS SQL 2012, MySQL 10g. |
| No-SQL : HBase, Cassandra. |
| Hadoop Distribution : Cloudera 5.4, Hortonworks 2.5. |
| Development tools : Eclipse 4.5, Intellij IDEA. |
| Operating Systems : UNIX, Windows XP/7/10, OSX (el caption). |
| Cloud : AWS (Amazon Web Services), Data bricks Cloud. |
| Cluster Mangers : Mesos, Yarn. |
| Software Development : AGILE, Scrum. |
| File Formats : JSON, Parquet, CSV, Sequence. |
| Tracker : Jira, Pivotal Tracker. |
| Data Modeling : Erwin. |
| Containers : Docker, Kubernieties. |
| BI Reporting Tools : Tableau. |
| Web Services : Restful, SOAP. |
| Authentication : Kerberos. |
| Serialization Techniques : Java Serialization, Kryo Serialization. |

**Professional Experience:**

**Project: DPEA (Distributed Portfolio Evaluation and Assessment) Jul 2016 – Till date**

**Client: DISCERN**

**Role: Hadoop/Spark Developer**

**Location: San Mateo, CA.**

**Description:** DPEA is a flexible and highly performant product. It provides Data Analysts to run various “what-if” analysis on individual or group of stocks. What used to take hours/days previously, with new data processing/analytics platform cut it down to minutes/hours. This is enabling users, primarily data scientists, to develop various predictive analytics models.

This system pulls information from multiple data sources and ingests it into the systems Data Lake. It helps in identifying trends – price moving up/down, whether the stock price is in alignment with Sector trends or deviating, Stock Volume Data Story, for profitable stock purchasing/selling.

**Responsibility**

* Involved in the overall architecture design for the system.
* Perform ELT jobs and automate ELT operations using Autosys or cronjobs.
* Writing Python Scripts for web scraping data form various stock market websites.
* Developing scripts using Scala for spark Streaming find out Share Market Trends.
* Developing Python scripts for data filtering and cleansing.
* Data Ingestion from multiple sources like RDBMS & Amazon S3 services and Weblogs.
* Perform SQOOP Incremental Import Job, Shell Script & CRONJOB for importing data into AWS S3.
* Imported data from RDBMS Tables to Hive using Sqoop.
* Created Hive partition and creating buckets to improve performance and optimize queries for low latency.
* Developed a Shell Script which dynamically downloads the Amazon S3 Data files into the HDFS system.
* Configuring AWS SNS for data ingestion into Flume Agent when data is added into MySQL. Writing MySQL triggers for the notification system.
* Configuring Amazon Kinesis pipelines to S3 storage for data ingestion.
* Created Datasets and Data frames for data transformation.
* Creating DStreams for processing Micro Batches in Spark Streaming.
* Implemented incremental import for S3 CSV files.
* Configuring Kyro Serialization for efficient Data Transmitting across the network.
* Using different SerDe techniques to improve performance.
* Handling Different Data Formats Like CSV, Json, Sequence File. Having knowledge of Avro, Parque File Formats.
* Perform experiments with Latest Version of Spark Features like (Structured Streaming, Blink DB)
* Migrating Data form HDFS to AWS S3 using DistCp.
* Used Agile development process and practices.

**Environment:** Kafka, Spark 2.0, Hive, Spark Streaming (Scala), Spark SQL Mesos, Cassandra.

**Client: Citi Group Aug 2015 - Jul 2016**

**Role: Hadoop Developer**

**Location: Denver, CO**

**Description:** Utilized the Hadoop cluster for staging & analyzing data from various data sources including web service API’s, relational databases and text files from a federal database containing safety related complaints, legal letters and other consumer product safety data.

**Responsibilities:**

* Developed Data pipeline using Sqoop, Flume to store data into HDFS and further processing through Spark.
* Creating Hive tables with periodic backups, writing complex Hive queries to run on spark.
* Involving in writing Map Reduce Programs to process the Data.
* Implemented partitioning, bucketing and worked on Hive, using file formats and compressions techniques with optimizations.
* Created Hive Generic UDF's to process business logic that varies based on policy.
* Experience in customizing map reduce framework at different levels like input formats, data types, custom SerDe and practitioners.
* Pushed the data to RDBMS Systems for mount location for Tableau to import it for reporting.
* Continuous monitoring and managing the Hadoop cluster using Cloudera Manager.
* Implemented Spark to migrate map reduce jobs into Spark RDD transformations, streaming data using Spark streaming.
* Developed Spark scripts using Scala, Spark SQL to access hive tables in spark for faster data processing.
* Configured build scripts for multi module projects with Maven.
* Automated the process of scheduling workflow using Oozie and Autosys.

**Environment**: Hadoop, Cloudera, HDFS, Hive, Spark, Sqoop, Flume, Java, Scala, Shell-script, Impala, Eclipse, MySQL.

**Client: Sears Aug 2014 - July 2015**

**Role: Hadoop Developer**

**Location: Chicago, IL**

**Description:** Using Hadoop To Target users who are searching for products or service based on search keywords and location display advertisement and product recommendations precisely. Advertisement based on the requirement stated by the clients. On Approval by the client, the advertisement is Published with Targets based on keyword tags.

**Responsibilities:**

* Using flume to gather data from source (user log data).
* Filter the data and ingest the data into appropriate schemas and tables to support the rule and analytics.
* Developed custom User Defined Function (UDF’s) in Hive to transform the large volumes of data with respect to business requirement.
* Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
* Responsible for building scalable distributed data solutions using Hadoop.
* Involved in loading data from edge node to HDFS using shell scripting.
* Developed Spark code using Scala and Spark-SQL for faster testing and processing of data.
* Implemented scripts for loading data from UNIX file system to HDFS.
* Implemented a script to transmit sysprin information from MySQL to hive and HBase.
* Experience in loading and transforming of large sets of structured data.
* Automated workflow using Shell Scripts.
* Good experience in Hive partitioning, bucketing and perform different types of joins on Hive tables.
* Experience in Hadoop 2.x with spark and Scala.
* Managed Hadoop jobs using Oozie workflow scheduler system for Map Reduce, Hive, Pig and Sqoop actions.
* Good knowledge on Data Ingestion and Data Processing.
* Used Spark SQL to process the huge amount of structured data.
* Experience in managing and reviewing Hadoop log files.
* Used Oozie workflow engine to run multiple Hive jobs.
* Exploring with the Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, Spark YARN.
* Analyzed large amounts of data sets to determine optimal way to aggregate and report on it.
* Responsible to manage the test data coming from different sources.
* Responsible for developing batch process using Unix Shell Scripting.

**Environment:** Apache Spark, Scala, Hadoop, HDFS, Hive, Sqoop, HBase, Unix, Kafka, Oozie, Cloudera CDH5x.

**Client: Norwich Union Life Oct 2012 – Jul 2014**

**Role: ETL Developer**

**Location: Chicago, IL**

**Description**: Norwich Union an AVIVA company, The largest insurance company in United Kingdom. The Finance Transformation project of Norwich Union will be responsible for the implementation of the new complex requirements of NUL and will also automate many sub routines which were manually calculated previously. The project will also be responsible for the accuracy of data in the Reports that will be run over the warehouse.

**Responsibilities:**

* Involved in the reverse engineering of the existing DCS system to get the ETL requirements
* Analyzed various Business Rules Engines available in the market and provided the list of features to the client
* Involved in the preparation of High level ETL architecture document.
* Creating ETL job to extract NUL legacy file systems.
* Data Cleansing, Noise reduction and improving Data quality.
* Store the data that extracted data into Data ware house staging Tables.
* Involved in the analysis of the Paymaster Source system.
* Involved in the Build of Informatica mappings.
* Responsible for creation of Unit test plans and System test plans
* Involved in creating and reviewing the ETL test cases for both Unit and System testing

**Environment:** Informatica 8.6, Teradata, Tableau, Unix, Windows NT, Oracle 9i, Sql Server, DB2.

**Client: IDRBT May 2009 - Aug 2012**

**Role: Software Developer**

**Location: Hyderabad, India.**

**Description**: Involving all phases of SDLC and gathering requirements and implementing the requirements testing and implementing change requests for production software. Issuing certificates for customer who are using our software

**Responsibilities:**

* Involved in Design, Development and Support phases of Software Development Life Cycle (SDLC)
* Reviewed the functional, design, source code and test specifications.
* Involved in developing the complete front end development using Java Script and CSS
* Author for Functional, Design and Test Specifications.
* Developed web components using JSP, Servlets and JDBC.
* Designed tables and indexes.
* Implementing change requests.
* Designed, Implemented, Tested and Deployed Enterprise Java Beans both Session and Entity using WebLogic as Application Server.
* Developed stored procedures, packages and database triggers to enforce data integrity. Performed data analysis and created crystal reports for user requirements.
* Implemented the presentation layer with HTML, XHTML and JavaScript.
* Implemented Backend, Configuration DAO, XML generation modules of DIS.
* Analyzed, designed and developed the component.
* Used JDBC/ODBC Connection for database access.
* Used Spring Framework for developing the application and used JDBC to map to Oracle database.
* Unit testing and rigorous integration testing of the whole application.
* Written and executed the Test Scripts using JUNIT.
* Developed XML parsing tool for regression testing.
* Prepared the Installation, Customer guide and Configuration document which were delivered to the customer along with the product.

**Environment:** Java, JavaScript, HTML, CSS, JDK 1.5.1, JDBC, Oracle10g, XML, XSL, Solaris and UML.

**Educational Summary:**

* Bachelor of Technology

**Certifications:**

* Apache Spark Certified Developer