|  |
| --- |
| **Venkata Pentapalli** |
| Mail: Venkata.3821@gmail.com |
| Phone: +1 469 846 8345 |

**Professional Summary:**

Having over **9 Years** of experience in Information Technology Industry which includes around 5 years of experience in **Big** **Data** **Ecosystem** like **Hadoop** and **Spark** **Ecosystems** and over 3 years of experience in Java in Developing, Implementing and maintenance of various web-based applications.

* Excellent knowledge on **Hadoop** **Architecture** and ecosystems such as **HDFS**, **Name Node, Data Node** and **Map** **Reduce** programming **paradigm**, Hive, **Sqoop**, **Kafka**, **HBase**, **Cassandra, MongoDB, Oozie, Zookeeper, Flume, Impala, Spark** with **Scala** and **PySpark**.
* Experience in analysing data using **Hive QL, HBase** and custom **Map** **Reduce** programs in Java.
* Experience in importing and exporting data using **Sqoop** from **HDFS** to Relational Database Systems and vice-versa.
* Experienced in creating data pipelines using the **Azure Synapse Analytics** and in **Databricks**.
* Good Exposure to **Azure Cloud, Azure Data Factory, ADLS Gen2, Azure Devops** (VSTS), portal services
* Worked with **Azure Databricks n**otebooks to validate the inbound/out bound from an external source like Amperity
* Expertise in implementing **Ad-hoc** queries using **Hive** **QL** and good knowledge in creating **Hive** tables and loading and analysing data using hive queries.
* Expertise in developing Hive Generic **UDF's** to implement complex business logic to incorporate into **Hive QL**.
* Developed **Apache** **Spark** jobs using **Scala** in test environment for faster data processing and used **Spark** **SQL** for **querying**.
* Experienced working with **Spark** **Streaming** and **Kafka** for real-time data processing.
* Used **Spark** **Data** Frames API over **Cloudera** platform to perform analytics on **Hive** data.
* Developed **Apache** **Spark** jobs using Scala in test environment for faster data processing.
* Worked on loading **CSV/TXT/AVRO/PARQUET** files using **Scala**/**Python** language in **Spark** **Framework** and process the data by creating **Spark** **Data** **frame** and **RDD** and save the file in parquet format in **HDFS**.
* Created dataflow between **SQL** **Server** and **Hadoop** **clusters** using **Apache Nifi.**
* Working knowledge of **Amazon's Elastic Cloud Compute** (**EC2**) infrastructure for computational tasks and Simple Storage Service (S3) as Storage mechanism.
* Working experience on **Hortonworks** distribution and **Cloudera** **Hadoop** distribution versions **CDH4** and CDH5 for executing the respective scripts.
* Good knowledge in working with **Impala** and **Kafka**.
* Experienced in moving data from different sources using **Kafka** producers, consumers and pre-process data using **Storm** **topologies**.
* Used **Oozie** and **Zookeeper** operational services for coordinating cluster and scheduling workflows.
* Good understanding and experience with Software Development methodologies like **Agile**, **Waterfall** and performed Testing such as **Unit**, **Regression**, **White-box, Black-box.**
* Monitor the **ETL** process job and validate the data loaded in Vertica/Teradata DW.
* Experience in **Web Services** using **XML**, **HTML**.
* Leveraged Google Cloud Platform Services (**GCP**)to process and manage the data from streaming and file-based sources.
* Experienced in running query using Impala and used **BI** tools to run ad-hoc queries directly on Hadoop.
* Administration of **Hadoop** and **Vertica** clusters for structured and unstructured data warehousing.
* Worked on version control tools like **CVS, GIT, SVN**.
* Well Experience in projects using **JIRA, Testing, Maven, MS Build** and **Jenkins** build tools.
* Hands on Experience in writing **SQL** and **PL/SQL** queries.

**Technical Skills:**

* **Big Data Ecosystem**: Hadoop, Map Reduce, Hive, YARN, Kafka, Flume, Sqoop, Impala, Oozie, Zookeeper, Pig, Spark, Ambari, MongoDB, Cassandra, Storm.
* **Hadoop Distributions**: Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapR and Apache
* **Cloud Technologies:** Azure SQL Database, Azure Data factory, Azure Synapse Analytics, AWS,GCP
* **Languages**: Java, Python, SQL, Scala, and JavaScript
* **No SQL Databases**: Cassandra, MongoDB and HBase
* **DB Languages:** MySQL, PL/SQL, PostgreSQL and Oracle
* **Java Technologies:** Servlets, JavaBeans, JSP, JDBC, JNDI, EJB and struts
* **Methodology:** Agile, waterfall
* **Development / Build Tools:** Eclipse, Ant, Maven, IntelliJ, JUNIT and log4J.
* **Third Party Tools**: Outline Extractor, SQL Developer, Putty, WINSCP.

**Educational Qualification:**

* **Bachelors:** Sunrise University Computer Science and Engineering **GPA:** 3.77
* **Maters :** University of central missouri Big Data Analytics

**Professional Experience:**

**Source InfoTech Inc.- Client: AT&T Big Data Inc. Dallas Jan 2021 – present**

**Role: Big Data Consultant**

**Description:** Project goal is to understand customer touchpoint across video and broadband. The end objective is to help a customer achieve better experience by reducing the number of calls calling to the care center. Create data pipelines and design a framework to create a centralized data set which is used for Journey analytics and BI.

**Responsibilities:**

* Design the data flow to import data (which may be structured, unstructured or semi-structured data) from multiple data sources like **Aster**, **Teradata**, **Vertica** and **SAS**. into the Hadoop data lake using SOAP Web Services, File Transfer Protocols, **Sqoop**, Map Reduce, **Hive** and **Pig**.
* Perform crucial transformations and query the loaded data using **Hive**, **SparkSQL** and build reporting tables.
* Coordinated with business customers to gather business requirements. And also interact with other technical peers to derive technical requirements.
* Perform Text Cleansing by applying various transformations using **Spark** Data frames and **RDDS**
* Gather business requirements and design and develop data ingestion layer and presentation layer.
* Highly motivated and versatile team player with the ability to work independently & adapt quickly to the environment.
* Experience on Migrating SQL database to **Azure data Lake, Azure data lake Analytics, Azure SQL Database**, NOSQL, **Databricks** and **Azure SQL Data warehouse** and controlling and granting database access and Migrating On premise databases to Azure Data Lake store using **Azure Data factory.**
* Performed ad-hoc queries on structured data using **Hive QL** and used **Partition, bucketing techniques** and joins with Hive for faster data access.
* Develop **Hive** queries on external tables to perform various analysis.
* Used HUE for running **Hive** queries. Created partitions according to data using Hive to improve performance.
* Importing and exporting data into HDFS and HIVE using **Sqoop**.
* Responsible for loading data from UNIX file systems to HDFS. Installed and configured Hive and written Hive UDFs.
* Developed **Spark** Applications by using Scala and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Used **Spark SQL** on data frames to access hive tables into spark for faster processing of data
* Designed and developed jobs to validate the data post migration such as reporting fields from source and designation systems using **Spark SQL RDDs** and Data Frames/Datasets.
* Used **Spark Data Frame** API to process Structured and Semi Structured files and load them into AWS **S3** Bucket.
* Used **Spark Data Frames** Operations to perform required Validations in the data and to perform analytics on the Hive data.
* Used Different **Spark** Modules like **Spark core**, **Spark SQL**, **Spark Streaming**, **Spark Data sets and Data frames.**
* Worked on Apache **Solr** for indexing and load balanced querying to search for specific data in larger datasets­­­
* Worked on Spark Streaming and Structured Spark streaming using **Apache Kafka** for real time data processing.
* Research on Azure **ADF** to migrate data from on premise datastores to Cloud.
* Responsible for developing multiple **Kafka** Producers and Consumers from scratch as per the software requirement specifications.
* Involved in reading uncompressed data formats like **Gzip**, **Avro**, **Parquet** and **compressed** the same according to the business logic by writing generic code.
* Extract Real time feed using **Kafka** and **Spark Streaming** and convert it to RDD and process data in the form of Data Frame and save the data as Parquet format in HDFS.
* Developed workflows using **Oozie** to automate the tasks of loading the data into HDFS and pre-processing with Pig.
* Used the JSON and XML SerDe's for serialization and de-serialization to load JSON and XML data into HIVE tables.
* Working experience on **Cloudera** **Hadoop distribution** version **CDH5** for executing the respective scripts.
* Worked on multiple clusters in managing the Data in HDFS for Data Analytics.
* Involved in Agile methodologies, daily Scrum meetings.

**Environment**: Hadoop, HDFS, Hive, Cassandra, Sqoop, Oozie, SQL, Kafka, Spark, Scala, AWS, GitHub, Azure, Big Data Integration, Impala, Ambari, Impala, Apache Solr, Gzip, Avro, Parquet, JSON, XML SerDe's.

**Client: Nike Inc Inc, Beaverton, OR Jan 2019 –Dec 2020**

**Role: Hadoop Developer**

**Project Description:** Nike, Inc. is an American multinational corporation that is engaged in the design, development, manufacturing and worldwide marketing and sales of footwear, apparel, equipment, accessories, and services. Consumer Knowledge is part of Consumer Digital Technologies (CDT). It exists to enable Direct to-Consumer (DTC) and Global Consumer Knowledge COE (GCK) data scientists and analysts with the platforms, tools, and data to deeply understand consumer behavior so that they can inform the strategy for Nike's consumer-facing digital products and experiences. The objective of the project is enhanced and expand NIKE's ability to gather, analyze and leverage insights on its consumers to deepen our understanding and deliver personalized experiences.

**Responsibilities:**

* Hands on experience in loading data from **UNIX** file system to HDFS. Also performed parallel transfer of data from landing zone to the **HDFS** file system using **DistCp**.
* Experienced on loading and transforming of large sets of structured and semi structured data from **HDFS** through Sqoop and placed in HDFS for further processing.
* Designed appropriate partitioning/bucketing schema to allow faster data retrieval during analysis using **Hive**.
* Involved in processing the data in the **Hive** tables using HQL high-performance, low latency queries.
* Transferred the analyzed data across relational database from **HDFS** using **Sqoop** enabling BI team to visualize analytics.
* Developed and executed a migration strategy to move Data Warehouse from an **Oracle** platform to **AWS Redshift**
* Configured **Spark** streaming to receive real time data from the **Kafka** and store the stream data into **AWS S3** using **Scala**
* Developed custom aggregate functions using **Spark SQL** and performed interactive querying.
* Managing and scheduling Jobs on a **Hadoop** cluster using **Airflow** DAG.
* Involved in creating **Hive** tables, loading data, and running **Hive** queries in those data.
* Extensive working knowledge of partitioned table, **UDFs**, performance tuning, compression related properties in **Hive**.
* Work with Data Engineering Platform team to plan and deploy new Hadoop Environments and expand existing **Hadoop clusters**.
* Deploy **Informatica** objects in production repository.
* Monitor and debug **Informatica** components in case of failure or performance issues.

**Environment**: Hadoop technologies (Spark, Hive, Impala, Sqoop), AWS, Informatica 9.1, Oracle, Autosys, Unix, Hdfs, DistCp, Airflow DAG, UDFs.

**Client: Macy’s , New York Oct 2017– Dec 2018**

**Role: Big Data Consultant**

**Project Description:** Real Time Machine Learning Platform uses to collects and analyzes large amounts of data from our customers 24×7 from several data points - websites, mobile apps, Macy's Credit card, social media, and coupons redemption. All these data are collected, aggregated, and analyzed in the Hadoop cluster to find shopping patterns make cross sell, up sell business decisions and devise targeted marketing strategies. RAMP main use case for semantics for recommendations for anonymous and logged in users in MACYS Inc / Bloomingdales. In all cases the engine will consider all that is known or inferred about the user and her actions to generate recommendations via correlations (co-occurrences).

**Responsibilities:**

* Designed and implemented **Distributed/Cloud Computing** (**Map Reduce/Hadoop, Pig, HBase, AVRO, and Zookeeper**), Installed and configured Hadoop **MapReduce**, **HDFS**, developed multiple **Map Reduce** jobs in java for data cleansing and pre-processing.
* Processed data into **HDFS** by developing solutions, analysed the batch data using **Map Reduce, Pig, Hive,** and produce summary results from Hadoop to downstream systems.
* Importing and exporting data into **HDFS** and **Hive** using **Sqoop**.
* Extracted files from **RDBMS** through **Sqoop** and placed in **HDFS** and processed.
* Worked on various **performance optimizations** like using distributed cache for small datasets, **Partition**, **Bucketing** in **hive** and **Map** Side joins and **Reducer** side joins.
* Replaced default Derby metadata storage system for **Hive** with MySQL system.
* Developed the **Pig UDF’s** to pre-process the data for analysis.
* Developed a custom File System plug in for **Hadoop** so it can access files on Data Platform. This plugin allows **Hadoop Map Reduce** programs, **HBase**, **Pig** and **Hive** to work unmodified and access files directly.
* Designed and implemented **Map reduce** based large scale parallel relation learning system.
* Used **UC4** workflow engine to run multiple Hive and Pig jobs which run independently with time and data availability.
* Involved monitoring **Autosys**'s file watcher jobs and testing data for each transaction and verified data weather it ran properly or not using UC4.

**Environment**: Java (JDK 1.6), Eclipse, Oracle 10g, Hadoop Distribution of Cloudera, Hive, Spark, HBase, Map Reduce, HDFS, Pig, Oracle 11g/10g, LINUX, UNIX Shell Scripting, AVRO, Zookeeper, Sqoop, RDBMS, Pig, UC4.

**Clint: Source One Solutions, Hyderabad, INDIA July 2014 – Sept 2017**

**Role: Java Developer**

**Project Description**: Hugh Solutions is an IT solutions company known for achieving business objectives and bottom-line results through the smart architecting, implementation, and management of technology.

**Responsibilities:**

* Extensive Involvement in Requirement Analyzes and system implementation.
* Actively involved in **SDLC** phases like Analyzes, Design and Development.
* Responsible for developing modules and assist in deployment as per the client’s requirements.
* Application is implemented using JSP and servlets are used for implementing Business logic.
* Developed utility and helper classes and Server-side Functionalities using servlets.
* Created **DAO** Classes and Written Various **SQL** **queries** to perform **DML** Operations on the data as per the requirements.
* Created Custom Exceptions and implemented Exception handling using Try, Catch and Finally Blocks.
* Developed user interface using **JSP**, **JavaScript** and **CSS** Technologies.
* Implemented User Session tracking in **JSP**.
* Involved in Designing **DB** Schema for the application.
* Implemented Complex **SQL Queries**, **Reusable** **Triggers**, **Functions**, **Stored** **procedures** using **PL/SQL**.
* Worked in pair programming, Code reviewing and debugging.
* Involved in Tool development, Testing and **Bug** **Fixing**.
* Performed **unit** **testing** for various modules.
* Involved in **UAT** and production deployments and support activities.

**Environment:** Java, J2EE, Servlets, JSP, SQL, PL/SQL, HTML, JavaScript, CSS, Eclipse, Oracle, MYSQL, IBM WebSphere, JIRA, PL/SQL.