**SRAVAN KUMAR**

**Sudheer.r@tekprosol.com**

**609 981 8255**

**Professional Summary:**

Above 14+ years’ experience as Data Architect/Data Developer in Hadoop Ecosystem using Spark, Scala, Hive, HBase, Pig and Sqoop for complex business problems, which involves real time effective analysis and processing of semi-structured and unstructured data.

* Experience in **Hadoop, Big Data technologies** such as **HDFS, MapReduce, Apache Pig, Impala, Hive, HBase**, **Sqoop, Flume, Oozie, Spark, Storm, Kafka, AWS, Amazon EMR, Zookeeper** and **Oracle**.
* Strong knowledge on implementation of **SPARK core** - **SPARK SQL and Spark streaming**
* Good Knowledge in **Amazon AWS** concepts like **EMR** and **EC2** web services which provides fast and efficient processing.
* Extensive experience in Big Data Analytics, **Hadoop, HDFS, MapReduce, YARN, Hadoop Ecosystem** and **Shell Scripting**.
* Experience in Using **Splunk dashboard** for monitoring real time traffic and analyzing traffic trend.
* Highly capable for processing large sets of **Structured, Semi-structured** and **Unstructured datasets** and **supporting Big Data applications**.
* Hands-on experience with Hadoop Ecosystem components like **Map Reduce (Processing), HDFS (Storage), YARN, Sqoop, Pig, Hive, HBase, Oozie, Zookeeper** and **Spark** for data storage and analysis.
* Good understanding of **Apache NiFi** and to automate data flow.
* Expertise in transferring data between a Hadoop ecosystem and structured data storage in a **RDBMS** such as **MY SQL, Oracle, Teradata** using **Sqoop**.
* Experience in **NoSQL** database **Hbase**.
* Experience in **Apache Spark** cluster and streams processing using **Spark Streaming**.
* Expertise in moving large amounts of log, streaming event data and Transactional data using **Flume**.
* Experience in developing **MapReduce jobs** in Java for data cleaning and preprocessing.
* Expertise in writing **Pig Latin, Hive Scripts** and extended their functionality using User Defined Functions (UDF's).
* Expertise in handling arrangement of data within certain limits (Data Layout) using Partitions and Bucketing in **Hive**.
* Expertise in preparing Interactive **Data Visualization using Splunk dashboard** from different sources.
* Hands-on experience in developing workflows execute **MapReduce, Sqoop, Pig, Hive and Shell Scripts** using **Oozie**.
* Experience working with **Cloudera Hue Interface** and **Impala**.
* Experience in **database architecture, design, development and maintenance**.
* Experienced in configuring **Data Guard on Linux** and Windows environments.
* Experience using Enterprise Manager Grid Control to monitor system performance
* Experience in Planning and Scheduling **Physical Backups** (offline/online) and Logical Backups with Export/Import data pump utilities
* Experience in migration of databases from one platform to another platform using transportable **tablespace**.
* Management of Schemas, Objects & Partitioning of **tables & indexes**.
* Monitoring of **CPU, Disk and Memory usage**, and File systems.
* Experience in Detection and diagnosing production system performance by analyzing Oracle trace and alert log files, **AWR, ADDM**, analyzing **SQL statements**, measuring system and database activity, Querying Oracle dynamic performance views.
* Experienced with tools/utilities: **OEM, RMAN**, Stats pack, **TOAD, SQL** Navigator, and SQL\* Loader, Data Pump (expdp/impdp).
* Experience in creating **db links** using listener.ora and tnsnames.ora files.
* Experience in process *I*mprovement, Normalization/Denormalization, Data extraction, cleansing and Manipulation.
* Converting requirement specification, Source system understanding into **Conceptual, Logical** and **Physical Data Model, Data flow (DFD).**
* Expertise in working with Transactional Databases like **Oracle, SQL server, My SQL,** and **Db2**.
* Expertise in developing **SQL queries**, Stored Procedures and excellent development experience with Agile Methodology.
* Good understanding of **NoSQL databases** and hands-on work experience in writing applications on NoSQL databases like **HBase, Cassandra and MongoDB**.
* Hands-on experience in Import/Export of data using **Hadoop Data Management tool SQOOP**.
* Proficient in configuring **Zookeeper**, **Cassandra & Flume** to the existing Hadoop cluster
* Written multiple **MapReduce** programs in **Python** for data extraction, transformation and aggregation from multiple file formats including **XML, JSON, CSV** and other compressed file formats.
* Experienced working **with JIRA** for project management, GIT for source code management,
* Strong Experience on writing **SQL Queries, PL/SQL, JPA** that includes Procedures, functions, triggers, cursors and packages.
* Ability to adapt to evolving technology, Strong sense of Responsibility and Accomplishment.
* Excellent leadership, interpersonal, problem solving and time management skills.
* Excellent communication skills both Written (documentation) and Verbal (presentation).

**Technical Skills:**

|  |  |
| --- | --- |
| Technology  | Hadoop Ecosystem, Oracle. |
| Big Data Ecosystem  | Hadoop, HDFS, MapReduce, Hive, Pig, HBase, Spark, Scala, Impala, Kafka, Hue, Sqoop, Oozie, Flume, Zookeeper, Cassandra, Cloudera CDH5, Python, PySpark, Solr and Hortonworks. |
| RDBMS/Databases  | Oracle 12c R1, 12c R2, 11g, 10g, 9i, MySQL, SQL Server, DB2, Mongo DB, Teradata, HBase, Cassandra. . |
| Methodologies  | Agile, WaterFall. |
| NOSQL Databases  | Cassandra, MongoDB, HBase. |
| Version Control Tools  | SVN, CVS, VSS, PVCS. |
| Reporting Tools:  | Crystal Reports, SQL Server Reporting Services and Data Reports, Business Intelligence and Reporting Tool (BIRT). |
| Tools | Oracle Enterprise Manager (OEM), TOAD, PUTTY, SQL Loader, SQL Developer, Data pump (expdp/impdp), AWR, STATSPACK, SQL Access Advisor, SQL Tuning Advisor  |
| Operating Systems | Red hat Linux, AIX, HP-UX, Windows 2000/NT/XP  |

**Professional Experience:**

**Client: Amazon, Remote Bay Area, CA Sep 2023 – Current**

**Role: Data Architect/Big Data Architect**

**Responsibilities:**

* Experience building and leading development of Data Integration and Data Modernization teams that design and deliver highly reliable and scalable systems.
* Dimension modelling for the DW for optimal storage, performance, scalability and easier programming of Analytical queries.
* Created Airflow DAGs for scheduling jobs and also altered DAGs to include jobs to be called inside them
* Created lifecycle policies for data in S3 buckets to save costs and optimal performance of the bandwidth.
* Integration of Data visualization tools like Tableau, Quick Sight and use of Athena and Query Editor
* Understanding the over-all Architecture of the system and ensure the deliverables are in-line with the proposed AWS architecture
* Build the AWS data processing solution focusing development of reusable components
* Work with AWS services like S3, Lambda, Step Functions any of the relational databases AWS - PrestoDB , RDS PostgreSQL
* AWS Lambda function development experience with Java and/or Python.
* Utilizing APIs and Kafka to streamline data processing and enhance system efficiency.
* Analyze data processing requirements, source data, and data domain models
* Author extract transform and load (ETL) scripts for moving and curating data into data sets for storage and use by a datalake, data warehouse and datamart.
* Monitoring MySQL database performance using tools like MySQL Enterprise Monitor, Percona Monitoring and Management, or custom scripts, and proactively tuning configuration parameters and resource utilization to optimize performance.
* Develop tools and procedures to monitor and automate system tasks on servers and clusters
* Collaborate with other teams to design, develop, and deploy data tools that support both operations and product use cases.
* Experience with AWS cloud and Big Data computing design, provisioning, and tuning.

**Environment**: Hadoop, HDFS, Hive, Airflow, UNIX Shell Scripting, Flume, Scala, Spark, Python, PySpark, Tableau

**Client: Asana, San Francisco, CA Dec 2021 – Sep 2023**

**Role: Big Data Developer**

**Responsibilities:**

* Responsible for Managing, Analyzing and Transforming terabytes of data and perform data validation while loading data into Cornerstone database.
* Responsible for Data Correction on already ingested data in Data Marts by developing spark-sql queries to extract and replace data on Hadoop cluster.
* Responsible for monitoring real time streaming jobs thru Splunk and take necessary action if Kafka jobs queued for longer duration based on its SLA.
* Responsible to resolve incidents raised by production support team by debugging failed files during ingestion and processing files back to Cornerstone database.
* Responsible for resolution of key issues upon escalation.
* Developing Python scripts to automate the cleaning process on the landing zone.
* Developing API end-points using Django application for internal applications.
* Experienced in using Kafka as a data pipeline between JMS (Producer) and Spark Streaming Application (Consumer).
* Experienced in creation of Hive tables and loading data incrementally into the tables using Dynamic Partitioning.
* Worked on SQL query optimization to increase performance of the job.
* Converted Teradata code to Spark SQL and performed query optimization.
* Migrating the needed data from Oracle, MySQL in to HDFS using Sqoop and importing various formats of flat files in to HDFS.
* Developed a shell script to create staging, landing tables with the same schema as the source and generate the properties which are used by Oozie jobs.
* Developed scripts to integrate Relational Database with NoSQL database (Hbase).
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD, Scala and Python.
* Worked with GIT CI/CD process to deploy code into production environment.
* Used Apache TEZ as an execution engine for building high performance batch and interactive data processing applications, coordinated by YARN on Hortonworks distributions.
* Worked using shell-scripting to deploy clusters and send data files as per the specifications given by label providers.
* Worked using JSON files for running the HIVE scripts and in retrieving the data after completion of the CI/CD pipelines.
* Worked on Hive by creating external and internal tables, loading it with data and writing Hive queries.
* Responsible for Debug, Optimization of Hive Scripts and also implementing Deduplication Logic in Hive using a Rank Key Function (UDF).
* Developed workflow in Oozie to automate the tasks of loading data into HDFS and pre-processing with Pig and Hive.
* Strong Knowledge of Apache NiFi which is used to automate data flow between systems.
* Responsible for Scheduling using Active Batch jobs and Cron jobs.
* Experienced in Jar builds that can be triggered by commits to GitHub using Jenkins.

**Environment**: Hadoop, HDFS, Map Reduce, Hive, HBase, Kafka, Zookeeper, Oozie, MapR, Oracle, Teradata SQL Server, UNIX Shell Scripting, Flume, Scala, Spark, Sqoop, Python.

**Client: Autodesk, San Rafael. CA Jan 2014 – Dec 2021**

**Role: Big Data Developer**

**Responsibilities:**

* Responsible for analyzing large data sets and derive customer usage patterns by developing new MapReduce programs.
* Experienced in creation of Hive tables and loading data incrementally into the tables using Dynamic Partitioning and Worked on Avro Files, JSON Records.
* Worked on Hive SQL for faster execution of Hive queries using AWS EMR.
* Worked with GIT CI/CD process to trigger the pipelines in order to generate DATA FILES.
* Developed HIVE scripts to extract data from DE Team S3 buckets.
* Developed Oozie workflow’s for executing Sqoop and Hive actions.
* Worked with HDFS commands to copy DATA from S3 bucket.
* Developed and executed HIVE queries on HUE browser.
* Used Apache TEZ has an execution engine for building high performance batch and interactive data processing applications, coordinated by YARN on Hortonworks distributions.
* Worked with LINUX OS in-order to write Test Cases.
* Have worked using AMAZON S3 services to store and retrieve data from S3 buckets.
* Developed and executed HIVE scripts in AWS ATHENA.

**Environment**: Hadoop, HDFS, Map Reduce, Hive, HBase, Kafka, Zookeeper, Oozie, MapR, Oracle, UNIX Shell Scripting, Flume, Scala, Spark, Sqoop, Python.

**Client: Logic Planet, NJ Aug 2011 – Jan 2014**

**Role: Sr. Hadoop/ Big Data Developer**

**Responsibilities:**

* Responsible for Writing MapReduce jobs to perform operations like copying data on HDFS and defining job flows on EC2 server, load and transform large sets of structured, semi-structured and unstructured data.
* Involved in generating the reports of the results of the scripts to analyze the necessities by using data visualization tool tableau.
* Designed and implemented custom writable, custom input formats, custom partitions and custom comparators in MapReduce.
* Developed pig scripts to transform the data into structured format and it is automated through Oozie coordinators.
* Developed Map reduce programs for the files generated by hive query processing to generate key, value pairs and upload the data to NoSQL database HBase.
* Responsible for troubleshooting issues in the execution of MapReduce jobs by inspecting and reviewing log files.
* Worked and used transformations, cleaning and filtering on imported data using Hive, Map Reduce, and loaded final data into HDFS
* Conducted data analysis with basic Python and wrangled data for data repositories.
* Configured Hive Meta store to use Oracle database to establish multiple user connections to hive tables.
* Developed a process for Sqooping data from multiple sources like SQL Server, Oracle and Teradata.
* Worked with NoSQL databases like HBase in creating HBase tables to load large sets of semi structured data coming from various sources.
* Performance optimizations on Spark/Scala. Diagnose and resolve performance issues.
* Responsible for developing Python wrapper scripts which will extract specific date range using Sqoop by passing custom properties required for the workflow.
* Developed scripts to run Oozie workflows, capture the logs of all jobs that run on cluster and create a metadata table which specifies the execution times of each job.
* Developed Hive scripts for performing transformation logic and loading the data from staging zone to final landing zone.
* Worked on Parquet File format to get a better storage and performance for publish tables.
* Involved in loading transactional data into HDFS using Flume for Fraud Analytics.
* Developed Python utility to validate HDFS tables with source tables.
* Designed and developed UDF’S to extend the functionality in both PIG and HIVE.
* Import and Export of data using Sqoop between MySQL to HDFS on regular basis.
* Responsible for developing multiple Kafka Producers and Consumers from scratch as per the software requirement specifications.
* Automated all the jobs for pulling data from FTP server to load data into Hive tables using Oozie workflows.
* Involved in developing Spark code using Scala and Spark-SQL for faster testing and processing of data and exploring of optimizing it using Spark Context, Spark-SQL, Pair RDD's, Spark YARN.

**Environment**: Hadoop, HDFS, Map Reduce, Hive, HBase, Kafka, Zookeeper, Oozie, Impala, Cloudera, Oracle, UNIX Shell Scripting, Flume, Scala, Spark, Sqoop, Python.

**Client: The New York Times. Manhattan, NY Sep 2010 – Dec 2011**

**Role: ETL Developer**

**Responsibilities:**

* Interacting with Client and Business Analysts to acquire Data requirements, Client requirements for the project.
* Extensively used Informatica Designer to create and manipulate source and target definitions, mappings, mapplets, transformations, Worklets, re-usable transformations, etc.
* Worked on Business Object's XI Reporting(Desktop Intelligence and Web Intelligence) as per my discussion with the Users
* Responsible for Leading the Offshore team (Informatica Developer/BO)
* Responsible for Migrating the Code from DEV to UAT to PROD
* Responsible for setting up the expectations to the Users and defining the time lines
* Working on creating as well as loading the data in to Star and Snowflake Schema (Fact and Dimensional Tables) by Slowly Changing Dimensions.
* Responsible for Leading the Offshore Team by guiding and assigning the task's to them.
* Worked on both Extracting and loading the data from and to Sales Force (Sales, Service & Support, Analytics, and their collaboration platform).
* Responsible for Creating the Solution design to define the Business Flow, Detail Design Document for the developers and preparing the Sub Doc's for the Code Migration from Dev to QA and QA to UAT and UAT to Prod.
* Enhanced the Dimensional Data Model (Star Schema) as per the Business Need's and worked on creating the Universe as well as generating the BO Reports (Web and Deski Reports).
* Worked on Unix Shell Script to pick the data files from a filestore and load them for further processing.
* Involved in Creating the Labels and Deployment Groups

**Environment**: SQL, PL/SQL, ORACLE, SQL Server, Informatica, ETL, UNIX Shell Scripting

**Education:**

* Bachelor’s in Electrical & Electronics Engineering, JNTU – India