VINAY KUMAR SURABHI
Data Engineer

(I am a Senior data engineer having 8+ years of experience in data engineering, Data Pipeline Designing and Development using Hadoop ecosystem such as HDFS, Hive, MapReduce, HBase and using programming languages Scala, Python. Currently I'm involved in designing and deploying the pipelines. I'm also having expertise in Hadoop, and spark. I have experience in working with different cloud services like AZURE and GCP, AWS.)

☎ **(845)261-1712**

in linkedin.com/in/vinay-kumar-761839262

✉ **Kv.vinay77@gmail.com**

SUMMARY:

- 8+ years of experience as Big Data Engineer /Data Engineer including designing, developing and implementation of data models for enterprise-level applications and systems.
- Experience in using cloud services Amazon Web Services (AWS) including EC2, 53, AWS Lambda and EMR, used Redshift for migration.
- Experience in creating complex data pipeline process using T-SQL scripts, SIS packages, Aptery workflow, PL/SQL scripts, Cloud REST APIs, Python scripts, GCP Composer, GCP dataflow.
- Experience in building ETL systems using python and in-memory computing framework (Apache Spark), scheduling and maintaining data pipelines at regular intervals in Apache Airflow.
- Experience in analyzing data using Spark SQL, HIVEQL, PIG Latin, Spark/Scala and custom Map Reduce programs in Java.
- Experience in Creating pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
- Experience of using C/CD techniques and processes DevOps/Git repository code promotion.
- Experienced with version control systems like Git, GitHub, to keep the versions and configurations of the code organized.
- Experience with structured (MySQL, Oracle SQL, PostgreSQL) and unstructured (NoSQL) databases. Strong understanding of relational databases.
- Familiar with cross platforms ETL using Python/JAVA SQL connector, PySpark Data Frame.
- Expertise in Data Extraction, Transforming and Loading (ETL) between different Systems using SQL tools (SSIS, DTS, Bulk Insert, and BCP).
- Experience in designing, modelling, performance tuning and analysis, implementing processes using ETL tool Pentaho Data Integration (PDI) tool for Data Extraction, transformation and loading processes. Designing end to end ETL processes to support reporting requirements.
- Designed aggregates, summary tables and materialized views for reporting.
- Extensive experience in installing and configuring Pentaho BI Server for ETL and reporting purposes.
- Experience in using Design Patterns such as MVC, Singleton and frameworks such as DJANGO.
- Knowledge on Google Cloud Platform (GCP) services like compute engine, cloud load balancing, cloud storage, cloud Data Proc, Cloud Pub/Sub, cloud SQL, Big Query, stack driver monitoring and cloud deployment manager. Experienced in version control tools like SVN, GitHub and CVS.
- Experienced working with JIRA for project management, GIT for source code management, JENKINS for continuous integration and Crucible for code reviews.

- Good experienced in Data Analysis as a Proficient in gathering business requirements and handling requirements management.
- Experience in migrating the data using Sqoop from HDFS and Hive to Relational Database System and vice-versa according to client's requirement.
- Experience with RDBMS like SQL Server, MySQL, Oracle and data warehouses like Teradata and Netezza. Proficient knowledge and hands on experience in writing shell scripts in Linux.

TECHNICAL SKILLS:

Languages	Shell scripting, SQL, PL/SQL, Python, R, PySpark, Pig, Hive QL, Scala, Regular Expressions
Hadoop Distribution	Cloudera CDH, Horton Works HDP, Apache, AWS
Big Data Ecosystem	HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Flume, Impala
Databases	Oracle 10g/11g/12c, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL, MS Access, Snowflake, NoSQL Database (HBase, MongoDB).
Cloud Technologies	Amazon Web Services (AWS), Microsoft Azure, GCP
Version Control	GIT, GIT HUB
IDE & Tools, Design	Eclipse, Visual Studio, Net Beans, Junit, CI/CD, SQL Developer, MySQL, SQL Developer, Workbench, Tableau
Operating Systems	Windows 98, 2000, XP, Windows 7,10, Mac OS, Unix, Linux
Data Engineer/Big Data Tools/Cloud/ETL/Visualization/Other Tools	Databricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Spring Boot, Flume, YARN, Hortonworks, Cloudera, MLlib, Oozie, Zookeeper, etc. AWS, Azure Databricks, Azure Data Explorer, Azure HDInsight, Linux, Bash Shell, Unix, etc., Tableau, Power BI, SAS, Crystal Reports, Dashboard Design

EDUCATION:

Bachelor’s in Computer Science Engineering,
Vel tech Rangarajan Dr.Sagunthala R&D institute of science and technology.

PROFESSIONAL EXPERIENCE:

Medicare (Baltimore, MD) **Apr 2021 – Present**
Role: Data Engineer

Responsibilities:

- Worked on complex SQL Queries, PL/SQL procedures and convert them to ETL tasks.
- Built data pipelines to move data from source to destination scheduling by Airflow.
- Developed BIX Extract application in Python to ingest Pega (Complaint System) files to HDFS and configure Airflow DAGs to orchestrate ETL workflow.
- Involved in Agile Development process (Scrum and Sprint planning).
- Involved in various sectors of business, with In-depth knowledge of SDLC (System Development Life Cycle) with all phases of Agile - Scrum, & Waterfall.
- Developed Map Reduce jobs using Java to process large data sets by fitting the problem into the Map Reduce programming paradigm.
- Developed Spark scripts by using Java, and Python shell commands as per the requirement.
- Worked with CI/CD tools such as Jenkins and version control tools Git, Bitbucket.
- Worked on source control tools like Tortoise SVN, CVS, IBM Clear Case, Perforce, and GIT.
- Created pipelines, data flows and complex data transformations and manipulations using ADF and PySpark with Databricks.
- Used the RUP and agile methodology to conduct new development and maintaining software.
- Developed central and local flume framework for loading large log files into the Data Lake.
- Designed and implemented distributed systems with Apache Spark and Python/Scala.
- Created Python / SQL scripts, to transform Databricks notebooks from Redshift table into Snowflake S3 buckets.
- Worked with Reporting developers to oversee the implementation of report/universe designs.
- Created visualizations of KPIs and critical financial metrics (Domo, Python).
- Worked on designing and implementing complex applications and distributed systems into public cloud infrastructure (AWS, GCP, Azure, etc.).
- Designed workflows using Airflow to automate the services developed for Change data capture.
- Created Power BI SSRS. Tableau and Domo reports based on the format specified in the design document.
- Built code for real time data ingestion using Java, MapR-Streams (Kafka) and STORM O Used Eclipse IDE to develop Spark java code to insert data to HBase.
- Responsible for creating a Data pipeline flows, scheduling jobs programmatically (DAG) in Airflow workflow engine, and providing support for the scheduled jobs.
- Worked with Informatica Cloud for data integration between Salesforce, Right Now, Eloqua; Web Services applications.
- Involved in modeling datasets from verity of data sources like Hadoop (using Pig, Hive, Spark), Teradata and Snowflakes for ad-hoc analysis and have fair understanding of AGILE methodology and practice.
- Generated SQL Scripts using python to extract Structured and non-structured data from various platforms - Teradata, Redshift, Snowflake, and Databricks.
- Built, maintained and tested infrastructure to aggregate critical business data into Google Cloud Platform (GP) Big Query and GCP Storage for analysis.
- Designed, implemented and owned administration of multiple public cloud environments (AWS & GCP). Worked with Jenkins CI for CI/CD and Git version control.
- Designed and developed data flow solutions using NIFI to transfer data to HDFS in Data Lake.

- Created numerous pipelines in Azure using Azure Data Factory v2 to get the data from disparate source systems by using different Azure Activities like Move &Transform, Copy, filter, for each, Databricks etc.
- Worked in automation, setup and administration of build and deployment CI/CD tools such as Jenkins, and integrated with Build Automation tools like ANT, Maven, Gradle, Bamboo, JIRA, Bit Bucket for building of deployable artifacts.
- Worked on all phases of data integration development lifecycle, real time/batch data pipelines design and implementation, and support of WU Digital Big Data ETL & Reporting track.
- Used to manage GitLab and Bit Bucket account for providing access to the Developers and storing the source code.
- Wrote SOL queries to identify and validate data inconsistencies in data warehouse against source system.

Environment: SOAP, REST APIs, SQL, Azure, ETL, APIs, cloud, UNIX, PL/SQL, CI/CD, Matplotlib, PyHive, Keras, Java, NoSQL- HBASE, Sqoop, Pig, MapReduce, Oozie, Spark MLlib

Fidelity Investments (Durham, NC) | Dec 2019 to Mar 2021
Role: Data Engineer

Responsibilities

- Designed & developed batch processing solutions by using Data Factory and Azure Databricks.
- Designed, developed and implemented solutions with data warehouse, ETL, data analysis and BI reporting technologies.
- Identified, evaluated, and documented potential data sources in support of project requirements within the assigned departments as per agile methodology.
- Created Python / SQL scripts, to transform Databricks notebooks from Redshift table into Snowflake S3 buckets.
- Extensively worked on Data Services for migrating data from one database to another database.
- Implemented various performance optimization techniques such as caching, Push-down memory-intensive operations to the database server, etc.
- Worked with developing customized UDF's in java to extend Hive and Pig Latin functionality.
- Involved in data from RDBMS and performed data transformations, and then export the transformed data to Cassandra as per the business requirement and used Cassandra through Java services.
- Involved in Agile development methodology active member in scrum meetings.
- Involved in continuous integration and deployment (CI/CD) using DevOps tools like Looper, Concord. Designed a workflow using Airflow to automate the jobs.
- Implemented a CI/CD pipeline with Jenkins, GitHub, Nexus, Maven and AWS AMI'S.
- Created several Databricks Spark jobs with Pyspark to perform several tables to table operations.
- Designed and Implement test environment on AWS.
- Created S3 buckets also managing policies for S3 buckets and Utilized S3 bucket and Glacier for storage and backup on AWS.
- Followed Agile & Scrum principles in developing. Involved in porting the existing on-premise Hive code migration to GCP (Google Cloud Platform) Big Query.
- Implemented both ETL and ELT architectures in Azure using Data Factory, Databricks, SQL DB and SQL Data warehouse. Built a data pipeline and data applications to analyze email marketing campaigns, using Power Shell, SQL Azure and Power BI.
- Built a dashboard using DOMO and Tableau to build various business and operational for Guest Emails to have better insights for the management.

- Supported current data processing and compliance initiative by creating technical and summary documentation.
- Participate in daily standups, bi-weekly scrums and PI panning. The New Management Services is a SAFe (Agile) certified organization.
- Involved in design and development of UI using ASP.Net after interacting with users for requirements. Transferred data from AWS S3 to AWS Redshift.
- Engineered PySpark report processing pipeline in AWS to lay a framework to migrate existing system off Cloudera and enable business users to create their own customer reports without support.
- Worked on developing Map-Reduce scripts in Python.

Environment: Python, AWS S3, AWS Redshift, AWS Data Pipeline, Spark, CI/CD, IBM DB2, Airflow, SAP ECC, SQL, Agile, ELT, S3, SOL DB, SQL AZURE, AWS.

AbbVie (Chicago, IL) | May 2018 – Nov 2019
Role: Data Engineer

Responsibilities:

- Participated in requirements sessions to gather requirements along with business analysts and product owners. Involved in Kafka and building use case relevant to our environment.
- Worked on implementation and maintenance of Cloudera Hadoop cluster.
- Pulled the data from data lake (HDFS) and massaging the data with various RDD transformations.
- Involved in building an information pipeline and performed analysis utilizing AWS stack (EMR, EC2, S3, RDS, Lambda, Glue, SQS, and Redshift).
- Responsible for developing data pipeline using flume, Sqoop and pig to extract the data from weblogs and store in HDFS.
- Developed Oozie workflow jobs to execute hive, Sqoop and MapReduce actions.
- Architected, Designed and Developed Business applications and Data marts for reporting.
- Imported the data from different sources like HDFS/HBase into Spark RDD and developed a data pipeline using Kafka and Storm to store data into HDFS.
- Collaborated with Business users for requirement gathering for building Tableau reports per business needs.
- Developed Pig Latin scripts for replacing the existing legacy process to the Hadoop and the data is fed to AWS S3.
- Developed continuous flow of data into HDFS from social feeds using Apache Storm Spouts and Bolts. Involved in loading data from UNIX file system to HDFS.
- Implemented the Big Data solution using Hadoop, hive and Informatica to pull/load the data into the HDFS system.
- Objective of this project is to build a data lake as a cloud-based solution in AWS using Apache Spark. Installed and configured Hadoop Ecosystem components.
- Developed Spark code using Scala for faster testing and processing of data.
- Apache Hadoop installation & configuration of multiple nodes on AWS EC2 system.
- Created Hive External tables to stage data and then move the data from Staging to main tables.
- Documented the requirements including the available code which should be implemented using Spark, Hive, HDFS, HBase and Elastic Search.

Environment: Hadoop, YARN, HDFS, Spark, flume, AWS, Sqoop, pig, MapReduce, UNIX, Zookeeper HBase, Kafka, Scala, NoSQL, Cassandra, Elastic Search, Sqoop

Progressive Corporation (Mayfield, OH) | Mar 2016 – Apr 2018
Role: Data Engineer

Responsibilities:

- Worked closely with Business Analysts to gather requirements and design a reliable and scalable data pipelines.
- Develop and add features to existing data analytic applications built with Spark and Hadoop on a Scala, Python development platform on the top of AWS services.
- Programming using Python, Scala along with Hadoop framework utilizing Cloudera Hadoop Ecosystem projects (HDFS, Spark, Sqoop, Hive, HBase, Oozie, Impala, Zookeeper etc.).
- Involved in developing spark applications using Scala, Python for Data transformations, cleansing as well as validation using Spark API.
- Developed several Accelerators/Tools as Spark/Step Functions for Workflow (UNIX Scripts as well) applications that saved a lot of Manual Efforts.
- Developed ETL data pipelines using PySpark on AWS EMR, also Configured EMR clusters on AWS.
- Provided end to end data solutions to business and analytics team ensuring end to end encryption by leveraging AWS cloud services and native python and shell scripting.
- Involved in trouble shooting spark jobs with the help of Spark UI and monitored spark jobs.
- Setup continuous integration/deployment of spark jobs to EMR clusters (used AWS SDK CLI).
- Integration of data storage solutions in spark – especially with AWS S3 object storage.
- Worked on all the Spark APIs, like RDD, Data frame, Data source and Dataset, to transform the data.
- Worked on both batch processing and streaming data Sources. Used Spark streaming and Kafka for the streaming data processing.
- Developed Spark Streaming script which consumes topics from distributed messaging source Kafka and periodically pushes batch of data to spark for real time processing.
- Built data pipelines for reporting, alerting, and data mining. Experienced with table design and data management using HDFS, Hive, Impala, Sqoop, MySQL, and Kafka.
- Developed python scripts for data cleaning, analysis and automating day to day activities.
- Developed shell scripts and SQL for data analysis and quality checks.
- Developed automated reports using Tableau, Python and MySQL to reduce the manual intervention.
- Experienced working with source formats, which includes - CSV, JSON, AVRO, JSON, Parquet, etc.
- Used AWS EMR clusters for creating Hadoop and spark clusters. These clusters are used for submitting and executing Scala and python applications in production.
- Responsible for developing data pipeline with Amazon AWS to extract the data from weblogs and store in HDFS.

Environment: Hadoop 2.7.7, HDFS 2.7.7, Apache Hive 2.3, Apache Kafka 0.8.2.X, Apache Spark 2.3, Spark-SQL, Spark-Streaming, Zookeeper, Pig, Oozie, Java 8,11, Python3, S3, EMR, EC2, Redshift, Cassandra, Nifi, Talend, HBase, Cloudera (CHD 5.X)

DataFactZ (Hyderabad, India) | May 2014 – Aug 2015
Role: Data Engineer

Responsibilities:

- Worked extensively along with business analysis team, scrum masters in gathering requirements and understanding the workflows of the organization.
- Involved in Data mapping specifications to create and execute detailed system test plans. The data mapping specifies what data will be extracted from an internal data warehouse, transformed and sent to an external entity.
- Analyzed business requirements, system requirements, data mapping requirement specifications, and responsible for documenting functional requirements and supplementary requirements in Quality Center.
- Wrote and executed unit, system, integration and UAT scripts in a data warehouse projects.

- Wrote and executed SQL queries to verify that data has been moved from transactional system to DSS, Data warehouse, data mart reporting system in accordance with requirements.
- Created the test environment for Staging area, loading the Staging area with data from multiple sources.
- Worked on data profiling and data validation to ensure the accuracy of the data between the warehouse and source systems.
- Monitored the Data quality of the daily processes and ensure integrity of data was maintained to ensure effective functioning of the departments.
- Developed data mapping documents for integration into a central model and depicting data flow across systems & maintain all files into electronic filing system.
- Worked and extracted data from various database sources like DB2, CSV, XML and Flat files into the Data Stage.
- Used and supported database applications and tools for extraction, transformation and analysis of raw data.
- Performed data analysis and data profiling using complex SQL on various sources systems including Oracle and DB2.
- Wrote SQL scripts to test the mappings and Developed Traceability Matrix of Business Requirements mapped to Test Scripts to ensure any Change Control in requirements leads to test case update.
- Involved in extensive data validation by writing several complex SQL queries and Involved in back-end testing and worked with data quality issues.
- Delivered file in various file formatting system (ex. Excel file, Tab delimited text, Coma separated text, Pipe delimited text etc.).
- Performed ad hoc analyses, as needed, with the ability to comprehend analysis as needed.

Environment: Oracle 9i, SQL, DB2, XML, ad hoc, Excel 2008, Data Validation