

**Avinash**

**Data Engineer/Data analyst**

avinash381.n@gmail.com

**Professional Summary:**

* 9+ years of experience as a big data engineer on SQL, SPARK, HIVE, ETL(Informatica), Data modeling (Erwin),Snowflake, HBase, and HDFS.
* over years of experience gained project management and problem solving skills
* Experienced in Data warehouse, Data modeling, Data integration, Data Migration using ETL Jobs.
* Delivery of innovative database/data warehousing solutions to Finance Industries with HA solutions.
* Involved in various projects related to Data Modeling, System/Data Analysis, Design and Development for both OLTP and Data warehousing environments.
* Sound knowledge on Normalization and Denormalization techniques on OLAP and OLTP systems.
* Acute knowledge on Spark architecture and real-time streaming.
* Experience with configuration of Hadoop Ecosystem components: Hadoop (Horton Works), HDFS,
* Experienced in troubleshooting of Data Stage jobs and addressing production issues like performance tuning and enhancement.
* MapReduce, Hive and good knowledge on Impala, Kafka, Apache Storm, Sqoop, Oozie, Hbase and Zookeeper.
* Good experience on test-driven development, complete application life cycle development.
* Designed and Developed Flink pipelines to consume streaming data from Kafka.
* Developed industry standard solutions with Data stage ETL jobs based on business requirements using various Data stage stages
* Expertise in mapping the business requirements into functional requirements.
* Working proficiency in code repositories like Gitlab.
* Ability to work in Agile (Scrum) methodology.

**Technical Skills:**

|  |  |
| --- | --- |
| **Programming Language:** | Scala, SQL |
| **Big Data Tools:** | Hadoop 3.1.1, Sqoop 1.4.7, MapReduce, Flink, pyspark, Spark, Spark SQL, Azure, Spark Streaming, Kafka, PySpark, ,Pig, Hive. |
| **Database:** | MYSQL ,Snowflake, PostgreSQL, Oracle DB, Hive , HBase |
| **Operating System:** | Linux kernel, Ubuntu, Amazon Linux |
| **System Software:** | WordPress, MS Office (Word/Excel/Power Point/ Visio/Outlook), Local List, Google analytics, zoom, slack, skype |
| **Data tool:** | Erwin, Informatica, Tableau |
| **Development**  **Methodologies:** | Agile/Scrum, UML, Design Patterns, Waterfall |
| **AWS tools:** | EC2, S3 Bucket, AMI, RDS Amazon Redshift. |

**Work Experience:**

**Application architect/ Data Engineer June 2022- Present**

**Client: Bank of America**

**Description :**

Detecting fraud transactions, worked with a team of 8 members and other teams as well.

**Role :**

* Creating autosys jobs and monitoring the functionality
* Unit tested the data between redshift and snowflake
* Developed data warehouse model in snowflake for over 100 datasets using whereScape
* Used spark and spark-SQL to read parquet data and create the tables in hive using the scala API
* Used flink streaming for pipelined flink engine to process data streams to deploy new API.
* Involved in implementing ELK infrastructure
* Designed and developed automation scripts and batch jobs to create data pipelines between multiple datasources, Spark based analytics platform (Databricks),
* Setup a GCP firewall rules in order to allow or deny traffic to and from the VM’s instance based on specified configuration.
* Worked on GCP services like compute engine, cloud load balacing, cloud storage and cloud SQL
* Using Kimball four step process (Business process definition, Grain declaration, Fact and Dimension identification), designed Dimensional Data Models for Loan Servicing and Loan Origination with daily transactional facts and Customer, Account, Loan status, Credit profile etc. slowly changing dimensions.
* Implemented spark using scala and utilizing Data frames and Spark SQL API for faster processing of data
* Managing and reviewing log files using Web UI and cloudera manager
* Designed and Developed Data stage Jobs to Extract data from heterogeneous sources, applied transform logics to extracted data and Loaded into Data Warehouse Databases.
* Experience in meeting expectations with hadoop clusters using Cloudera (CDH3 & CDH4) and Horton works.
* Troubleshooting build issue that occurred because of ELK work

**Environment**: Hadoop, HDFS, Sqoop, Hive, Cosmos, Spark 2.x, Pyspark, Apache Flink, Spark SQL, Spark, HBase, Kafka, Python, Scala, Teradata, Databricks, Teradata Sql, Bitbucket, Ansible

**Data Engineer/Azure**

**Client: Truist Bank, Raleigh, NC June 2021- March 2022 Description :**

Client journey project is about providing APR to clients based on their debit and credit annual transactions. Day 0 and Day 1 are incremental data load into our target data source. They provided home loan products and services, including refinancing, reverse mortgages, and loan modification solutions.

**Role :**

* Converting SQL queries to Spark DF for processing and loading data into target sources.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Utilized Spark Dataframe and Spark SQL API extensively for processing.
* Developed a shell script to extract all the databases DDL on the prem hive and create them on the AWS Data lake storage.
* Involved in converting HiveQL/SQL queries into Spark transformations using Spark DF in Scala.
* Designed and configured Azure virtual network, subnets, azure network setting, DHCP address blocks, DNS setting and security policies
* Worked on azure fabric, microservices, IOT & docker containers in and involved in setting up terraform continuous build integration system.
* Used azure internal load balancer to provide high avalibility for Iaas VMs & PaaS role instances.
* Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL) and processing the data in In Azure Databricks.
* Provided key inputs for deciding partition strategy for huge size tables of around 5 million records in Hive
* Involved in ESP job scheduling and rising RITM in service now.
* Involved in testing phase for data ingestion in Informatica powercenter
* Designed the Column families in Cassandra.
* Ingested data from RDBMS and performed data transformations, and then export the transformed data to Cassandra as per the business requirement.
* Developed Spark code to use Scala and Spark-SQL for faster processing and testing.
* Used Spark API over Hadoop YARN as execution engine for data analytics using Hive.
* This project was focused on customer clustering. Used the ETL Data Stage Director to schedule and running the jobs, testing and debugging its components & monitoring performance statistics.
* Created various kinds of reports using Power BI and Tableau based on the client's needs.
* Setting up Kerberos principals and testing HDFS, Hive, Pig, and MapReduce access for the new users.
* Migrated the computational code in hql to PySpark.
* Worked with Spark Ecosystem using Scala and Hive Queries on different data formats like Text file and parquet.
* Worked in migrating Hive QL into Impala to minimize query response time.
* Responsible for migrating the code base from Hortonworks Platform to Amazon EMR and evaluated Amazon eco systems components like Redshift.
* Collected the logs data from web servers and integrated in to HDFS using Flume
* Worked with NoSQL databases like Hbase in creating Hbase tables to load large sets of semi structured data coming from various sources.
* Developed Python scripts to clean the raw data.
* Developed Hive scripts in Hive QL to de-normalize and aggregate the data.
* Developed customized UDF's in java to extend Hive and Pig functionality.
* Extracted the data from Teradata into HDFS/Databases/Dashboards using SPARK STREAMING.
* Loaded Golden collection to Apache Solr using Morphline code for Business team.
* Worked on different file formats (ORCFILE, Parquet, Avro) and different Compression Codecs (GZIP, SNAPPY, LZO).
* Created applications using Kafka, which monitors consumer lag within Apache Kafka clusters.
* Involved testing in APS Data Loading, Data Seeding & Data Bridging strategy
* Using Spark-Streaming APIs to perform transformations and actions on the fly for building the common learner data model which gets the data from Kafka in near real time and Persists into Cassandra.
* Imported data from AWS S3 into Spark RDD, Performed transformations and actions on RDD's
* Maintained Hadoop Cluster on AWS EMR. Used AWS services like EC2 and S3 for small data sets processing and storage
* Design and document REST/HTTP, SOAP APIs, including JSON data formats and API versioning strategy.
* Used Apache Kafka for collecting, aggregating, and moving large amounts of data from application servers.
* Used Hibernate ORM framework with spring framework for data persistence and transaction management.
* Used MLlib framework in Spark streaming for auto suggestions on predictive intelligence and maintenance.
* Developed Python code to gather the data from HBase (Cornerstone) and designs the solution to implement using PySpark.
* Performance analysis of Spark streaming and batch jobs by using Spark tuning parameters.
* Worked along with the Hadoop Operations team in Hadoop cluster planning, installation, maintenance, monitoring and upgrades.
* Used micro services for data visualization and the functional challenges of planning and implementing some solutions.
* Implemented Nifi flow topologies to perform cleansing operations before moving data into HDFS.
* Started using Apache NiFi to copy the data from the local file system to HDP.
* Used File System check (FSCK) to check the health of files in HDFS.Used Amazon Cloud Watch to monitor and track resources on AWS
* Scheduled the ETL Jobs in AWS Glue developed through using lambda logics, (boto3), S3 to load into DynamoDB and Redshift.
* Designed a data analysis pipeline in Python, using Amazon Web Services such as S3, EC2 and Elastic MapReduce
* Worked in an Agile development environment in sprint cycles of two weeks by dividing and organizing tasks. Participated in daily scrum and other design related meetings.

**Environment**: Azure Databricks, Azure MLFlow (Model Serving, Model Enabling), Python, Pandas, PySpark, Azure Data Lake Storage, Azure SQL Server, Azure Blog Storage, Scikit Learn, Azure ML Cluster, Databricks Notebooks, GitLab, Rest API, Azure Data Factory, Azure SQL Datawarehouse.

**Data Engineer,**

**Client: BNY Mellon, New York, NY April 2020 - April 2021**

**Description** :

It’s a data migration project merging two databases into a single Lonarc 2.0 database.

**Role:**

* Data profiling and generating reports for the missing data and inconsistent data.
* Skilled in High-Level Design of ETL DTS Package for integrating data from heterogeneous sources (Excel, CSV, Oracle, MySQL, flat file, Text Format Data).
* Extensively involved in the Physical/logical modeling and development of Reporting Data Warehousing System.
* Working with DBA to design reports for DB replica latency trends, analyzing the transaction logs to find the root cause of the issues.
* Worked on Transactional logs to process them using Spark and saving them on required formats by applying various ETL tasks on log data and saving the data.
* Worked on Data ingestion to Kafka and Processing and storing the data Using Spark Streaming.
* Involved in tuning of Cassandra cluster by changing the parameters of Read operation, Compaction, Memory Cache, Row Cache.
* Installed and configured Apache Hadoop and Hive/Pig Ecosystems.
* Created MapReduce Jobs using Hive/Pig Queries.
* Developed the Pig UDF’S to pre-process the data for analysis.
* Worked on No Sql database Hbase for storing computed results.
* Worked extensively on search engines Elasticsearch, Novus (In house).
* Worked on workflow scheduling using Oozie.
* Developed MapReduce/Spark Python modules for machine learning & predictive analytics in Hadoop on AWS. Implemented a Python-based distributed random forest via Python streaming.
* Worked on Continuous Integration and Automation Testing Job scheduling using Jenkins and TFS.
* Analyzing Audit logs using Splunk, querying and designing views and dashboard on Splunk.
* Production Support by handling production bugs by reproducing in lower environments and fixing and moving them to prod environments by creating Hotfixes.
* Designed and created Solr Schemas to create Solr Collections.
* Laid the guidelines for improving the code quality by implementing TDD and developed an integrated test framework using JUnit, Mockito.
* Installed and Configured Hadoop cluster using AWS for POC purposes.
* Implemented CI/CD pipeline using Maven & Jenkins.
* Worked with CMDB teams on deploying builds to various environments.
* Involved in gathering requirements from different teams to design the ETL migration process from Existing Oracle and SQL server.
* Created conceptual, logical & physical data models with Erwin and did design review meetings with other team members to finalize the model.
* Created transformations that involve configuring the following steps: Table input, Table output, Text file output, CSV file input, Insert/Update, add constants, Filter, Value Mapper, lookup, join rows, merge join, sort rows, Database Lookup, Set Environment Variables.
* Created and developed Slowly Changing Dimensions tables SCD2, SCD3 to facilitate maintenance of history.
* Work in Development, QA, Pre prod, and Production Environments
* Working with the SCRUM team in delivering agreed user stories on time for every sprint.

**Citi Bank, Tampa, FL Jun 2017 – March 2020**

**Spark Developer**

Responsibilities:

* Worked on installing Kafka on Virtual Machine and created topics for different users
* Actively involved in designing the Hadoop ecosystem pipeline.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Involved in designing Kafka for multi data center cluster and monitoring it.
* Responsible for importing real time data to pull the data from sources to Kafka clusters.
* Worked with spark techniques like refreshing the table and handling parallelly and modifying the spark defaults for performance tuning.
* Implemented Spark RDD transformations to Map business analysis and apply actions on top of transformations.
* Involved in migrating MapReduce jobs into Spark jobs and used SparkSQL and Data frames API to load structured data into Spark clusters.
* Involved in using Spark API over Hadoop YARN as execution engine for data analytics using Hive and submitted the data to BI team for generating reports, after the processing and analyzing of data in Spark SQL.
* Performed SQL Joins among Hive tables to get input for Spark batch process.
* Worked with a data science team to build statistical model with Spark MLLIB and Pyspark.
* Involved in performing importing data from various sources to the Cassandra cluster using Sqoop.
* Worked on creating data models for Cassandra from Existing Oracle data model.
* Designed Column families in Cassandra and Ingested data from RDBMS, performed data transformations, and then exported the transformed data to Cassandra as per the business requirement.
* Used Sqoop to import functionality for loading Historical data present in RDBMS to HDFS
* Designed workflows and coordinators in Oozie to automate and parallelize Hive jobs on Apache Hadoop environment by Hortonworks (HDP 2.2)
* Configured Hive bolts and written data to hive in Hortonworks as a part of POC.
* Implemented ELK (Elastic Search, Log stash, Kibana) stack to collect and analyze the logs produced by the spark cluster.
* Developed Python script for start a job and end a job smoothly for a UC4 workflow
* Developed Oozie workflow for scheduling & orchestrating the ETL process.
* Created Data Pipelines as per the business requirements and scheduled it using Oozie Coordinators.
* Wrote Python scripts to parse XML documents and load the data in the database.
* Worked extensively on Apache Nifi to build Nifi flows for the existing Oozie jobs to get the incremental load, full load and semi structured data and to get data from Rest API into Hadoop and automate all the Nifi flows runs incrementally.
* Created Nifi flows to trigger spark jobs and used put email processors to get notifications if there are any failures.
* Developed shell scripts to periodically perform incremental import of data from third party API to Amazon AWS
* Worked extensively with importing metadata into Hive using Scala and migrated existing tables and applications to work on Hive and AWS cloud.
* Developed the batch scripts to fetch the data from AWS S3 storage and do required transformations in Scala using Spark framework.
* Used version control tools like GITHUB to share the code snippet among the team members.
* Involved in daily SCRUM meetings to discuss the development/progress and was active in making scrum meetings more productive.

**Environment:** Hadoop, HDFS, Hive, Python, Hbase, Nifi, Spark, MYSQL, Oracle 12c, Linux, Hortonworks, Oozie, MapReduce, Sqoop, Shell Scripting, Apache Kafka, Scala, AWS.

**Nielsen, Chennai, india Aug 2013 – May 2017 Hadoop Developer**

Responsibilities:

* Analyzing Functional Specifications Based on Project Requirement.
* Ingested data from various data sources into Hadoop HDFS/Hive Tables using SQOOP, Flume, Kafka.
* Extended Hive core functionality by writing custom UDFs using Java.
* Developing Hive Queries for the user requirement.
* Worked on multiple POCs in Implementing Data Lake for Multiple Data Sources ranging from Team Center, SAP, Workday, Machine logs.
* Developed Spark code using Scala and Spark-SQL/Streaming for faster testing and processing of data.
* Worked on MS Sql Server PDW migration for MSBI warehouse.
* Planning, scheduling and implementing Oracle to MS SQL server migrations for AMAT in house applications and tools.
* Worked on Solr Search Engine to index incident reports data and developed dashboards in Banana Reporting tool.
* Integrated Tableau with Hadoop data source for building a dashboard to provide various insights on sales of the organization.
* Worked on Spark in building BI reports using Tableau. Tableau was integrated with Spark using Spark-SQL.
* Developed Spark jobs using Scala and Python on top of Yarn/MRv2 for interactive and Batch Analysis.
* Created multi-node Hadoop and Spark clusters in AWS instances to generate terabytes of data and stored it in AWS HDFS.
* Developed work flows in Live Compare to Analyze SAP Data and Reporting.
* Worked on Java development and support and tools support for in house applications.
* Participated in daily scrum meetings and iterative development.
* Search functionality for searching through millions of files of logistics groups.