**Avinash S**

**A blue hexagon with white text

Description automatically generated** A blue and grey logo with white stars

Description automatically generated **A hexagon with white text

Description automatically generated**

**Sr. Data Engineer**

**Contact No:**  (980) 737-2407

**Email:** [**avinashsam789@gmail.com**](mailto:avinashsam789@gmail.com)

**LinkedIn:** [**https://www.linkedin.com/in/avinash-sam/**](https://www.linkedin.com/in/avinash-sam/)

**PROFESSIONAL SUMMARY:**

* With over 10+ years of IT expertise, specialized in end-to-end project design and development spanning sectors like **Insurance, Healthcare, Finance, Sales, and Marketing**. As a **Senior Data Engineer**, my role encompasses the entire Software Development Life Cycle **(SDLC)**, including User Interaction, Business Analysis/Modeling, Development, Integration, and Build Planning and Management.
* Solid experience in Dimensional Data modeling, Star Schema/Snowflake modeling, FACT & Dimensions tables, Physical & logical data modeling, **ERWIN 3.x, Oracle Designer**, **Data Integrator**.
* Proficient in developing **Spark** applications within **Databricks**, utilizing **RDD, Spark SQL, Data Frames**, and Spark Streaming for data extraction, transformation, and aggregation across multiple file-formats, uncovering valuable insights into customer usage patterns.
* Hands on experience on Unified Data Analytics with **Databricks**, Databricks Workspace User Interface, Managing **Databricks Notebooks**, Delta Lake with Python, Delta Lake with **Spark SQL.**
* Skillfully Extracting, Transforming, and Loading data from diverse sources into Blob storage, Cosmos DB, **Azure SQL**, and **Azure DW** (Synapse) using Azure Data Factory, **T-SQL, and U-SQL**.
* Good understanding of Spark Architecture with **Databricks**, Structured Streaming. Setting Up AWS and Microsoft Azure with **Databricks**, **Databricks Workspace** for Business Analytics, Manage Clusters in **Databricks**, Managing the Machine Learning Lifecycle.
* Experience in **GCP** **Dataproc**, **GCS**, Cloud functions, **BigQuery**, **Azure** **Data** **Factory** **DataBricks**.
* Strong Experience in working with **ETL Informatica** (10.4/10.1/9.6/8.6/7.1.3 ) which includes components **Informatica PowerCenter Designer**, Workflow manager, components **Informatica Intelligent Cloud Services(IICS)** Workflow monitor, **Informatica server** and Repository Manager.
* Designed and implemented the **Enterprise Data Warehouse** and Analytic project from scratch utilizing cloud-based infrastructure and the **Data Vault 2.0** design pattern to ensure agility and reliability.
* Experience in **Data Warehousing/ETL Architecture** designing and **Data Modeling** (**Star** & **Snowflake** Schema and **LDM** & **PDM**) for large data warehousing projects.
* Experience designing **Star schema** and **Snowflake** dimensional modeling using **Erwin tool**.
* Hands-on experience on AWS components such as **EMR**, **EC2**, **S3**, **RDS**, **IAM**, Auto Scaling, **CloudWatch**, SNS, **Athena**, Glue, Kinesis, **Lambda**, **Redshift**, DynamoDB to ensure a secure zone for an organization in AWS public cloud.
* Experience of using **CI/CD** techniques and processes DevOps/Git repository code promotion.
* Designed, developed and Implemented **ETL** Process using **IICS** Data Integration.
* Experience in configuring **Oracle Multitenant** Container Databases (CDB) and Pluggable Databases (PDB) in **Oracle 19c/12 RAC**.
* Highly knowledgeable in developing data marts in big data world in **BigQuery** or on - premises **Hadoop** clusters.
* Skilled at identifying the right cloud native technology for developing and maintaining big data flows in organizations.
* Experience with building data pipelines in **Python**/ **Pyspark**/ **HiveSQL**/**Presto**/**BigQuery** and building **python** DAG in Apache Airflow.
* Expertise in Creating, **Debugging**, **Scheduling** and **Monitoring** jobs using Composer **Airflow**.
* Strong knowledge in data preparation, **data** **modelling** and **data** **visualization** using **Power** **BI** and had experience in developing various analysis services using **DAX queries**.
* Hands on experience with different programming languages such as **Python**, **R**, **SAS**.
* Very keen in knowing newer techno stack that **Google** **Cloud** **platform** adds.
* Experience in using different Hadoop eco system components such as **HDFS**, **YARN**, **MapReduce**, **Spark**, **Pig**, **Sqoop**, **Hive**, **Impala**, and **HBase**, **Kafka**, and Crontab tools.
* Experience in developing **ETL** applications on large volumes of data using different tools: **MapReduce**, **Spark**-**Scala**, **Spark**, **SparkSql**, and **Pig**.
* Experience in using SQOOP for importing and exporting data from **RDBMS** to **HDFS** and Hive.
* Experience with **Unix**/**Linux** systems with bash scripting experience and building data pipelines.
* Strong **SQL** development skills including writing **Stored** **Procedures**, **Triggers**, **Views**, and User Defined functions.
* Expertise in designing and deployment of **Hadoop** cluster and different **Big** **Data** analytic tools including **Pig**, **Hive**, **Sqoop**, **Apache** **Spark** with Cloudera Distribution.
* Knowledge in various file formats in **HDFS** like avro, orc, parquet.
* Experience in building and architecting multiple Data pipelines, end to end **ETL** and **ELT** process for Data ingestion and transformation in **GCP** and coordinate task among the team.
* Hands on experience in using **Google** **cloud** **platform** for **bigquery**, **cloud** **dataproc** and **apache** **airflow** services.
* Have good Programming experience with **Python** and **Scala**.
* Hands in experience on **NoSQL** database like **Hbase**, **Cassandra**.
* Experience in using stackdriver service/ dataproc clusters in **GCP** for accessing logs for debugging.
* Experience with scripting languages like **PowerShell**, **Perl**, **Shell**, etc.
* Expert knowledge and experience in fact dimensional modeling (Star schema, Snowflake schema), transactional modeling and SCD (Slowly changing dimension)
* Expertise knowledge in **Hive** **SQL**, **Presto** **SQL** and **Spark** **SQL** for **ETL** jobs and using the right technology for the job to get done.
* Experienced in using Hortonworks on-premise Hadoop distribution and have written **Hive** **SQL** and **Sqoop** scripts.
* Experience in **GCP** **Dataproc**, **GCS**, Cloud functions, **BigQuery**, **Azure** **Data** **Factory** **DataBricks**.
* Strong Experience in working with **ETL Informatica** which includes components Informatica PowerCenter Designer, Workflow manager, Workflow monitor, Informatica server and Repository Manager.
* Experience in building efficient pipelines for moving data between **GCP** and **Azure** using Azure Data Factory.
* Experience in building power bi reports on Azure Analysis services for better performance when comparing that to direct query using **GCP** **BigQuery**.
* Extensive use of cloud shell **SDK** in **GCP** to configure/deploy the services **Data** **Proc**, **Storage**, and **BigQuery**.
* Created and wrote **shell scripts (Bash), Ruby, Python** and **PowerShell** for automating tasks.
* Extensive experience using **MAVEN, ANT** as build tools for the building of deployable artifacts from source code.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| Big Data Technologies | Hadoop, Apache Spark, Apache Kafka, Apache Cassandra, Apache HBase, Apache Hive, Apache Pig, Apache Storm, MapReduce, HDFS, Druid, Sqoop, Oozie, Flume, Apache Flink, NiFi, Zookeeper, Yarn. |
| Data Warehouses | SnowFlake, Big Query, Kudu Impala, Hive, RedShift, Snowflake, Teradata, Azure Synapse, Oracle, MySQL, PostgreSQL, TeradataR15/R14, Cosmos, Spark-Redis, Microsoft SQL Server, IBM DB2, Azure SQL Database. |
| Operating Systems | Windows, Linux (Centos, Ubuntu),MacOS. |
| Programming Languages | Python, Scala, Java, JavaScript, PySpark, HiveQL, R, SQL, PL/SQL, Pig Latin, HiveQL, Unix, Shell Scripting, C, Groovy. |
| NoSQL Databases | HBase, Cassandra, MongoDB, DynamoDB, MarkLogicS. |
| IDE Tools & Utilities | Idea, Eclipse, Pycharm, Aginity Workbench, Git |
| ETL | Informatica , Data stage9.1/11.5(Designer/Monitor/Director) |
| Job Scheduler | Control-M, IBM Symphony Platform, AMBARI, Apache Air flow |
| Reporting Tools | Tableau, Lumira,powerBI |
| Cloud Environment | **AWS** – S3, Amazon Glue, Kinesis, EMR, Step Functions, Lambda Functions, IAM. EC2, Glacier, Redshift, RDS, Lambda, CloudWatch, Kinesis, CloudFront, Athena, Quick Sight  **GCP** –Big Query, Cloud DataProc, GCS Bucket, G-Cloud Function, Apache Beam, Cloud Shell, Cloud Data Flow. |
| Machine-Learning Techniques | Linear and logistic Regression, Classification and Regression Trees, Random Forest, Associative rules, NLP and Clustering |
| Data Modeling Tools | Erwin r7.1/7.2, ER Studio V8.0.1 and Oracle Designer,Oracle 9i,12c,18c19c |
| Scrum Methodologies | Agile, Asana, Jira |
| Others | MS Office, RTC, Service Now, OPTIM, IGC(Info sphere Governance catalog), WinSCP, MSVisio |

**PROFESSIONAL EXPERIENCE**

**Barclays, New York city  Feb 2022 to Till Date**

**Sr. Data Engineer**

**Responsibilities:**

* Configured Spark streaming to get ongoing information from **Kafka** and stored the stream information to **HDFS** and **HBase**.
* Worked on development of data ingestion pipelines using **ETL tool, Talend & bash scripting** with big data technologies including but not limited to **Hive, Impala, Spark, Kafka,** and **Talend**.
* Creating **Databricks** notebooks using **SQL**, **Python** and automated notebooks using jobs.
* Responsible for Metadata Management, keeping up to date centralized metadata repositories using **Erwin modeling tools**.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark **databricks** cluster.
* Designed and implemented the **Enterprise Data Warehouse** and Analytic project from scratch utilizing cloud-based infrastructure and the **Data Vault 2.0** design pattern to ensure agility and reliability.
* Extensively working on Data Modelling tools **Erwin Data Modeler** to design the data models.
* Designed, developed and Implemented **ETL** Process using **IICS** Data Integration.
* Develop **HP Vertica SQL** scripts for bulk loading, delta loading stage & target tables Using Cloud Data Integration.
* Expertise in **Power BI** development with a proven track record in Data Warehouse (DWH) and Data Analytics.
* **RDDs** and **Scala** and involved in using **SQOOP** for importing and exporting data between **RDBMS and HDFS.**
* Developed data pipeline using **Flume**, **Sqoop**, **Pig** and **Java MapReduce** to ingest customer behavioral data into **HDFS** for analysis.
* Experience in working with **Flume** and **NiFi** for loading log files into **Hadoop**.
* Used **Python** and **Django** creating graphics, **XML processing**, data exchange and business logic implementation.
* Used **Informatica Power Center** for extraction, transformation, and loading (ETL) of data in the data warehouse.
* Working with **CI/CD** tools such as Jenkins and version control tools Git,Bitbucket
* Developed Spark Applications by using **kafka** and Implemented **Apache Spark** data processing project to handle data from various **RDBMS** and Streaming sources.
* Deployed the scripts on **AWS EMR** instance and made performance optimization changes. Code base has been maintained through **AWS Code Pipeline** and the scripts are scheduled through **AWS CloudWatch.**
* Developed Python based **Flask API** and used **AWS API Gateway** to provide access to internal team and used AWS **Lambda** for updating the metadata table after API calls are made.
* Build **AWS Glue** to automatic insertion of files coming from Vendor SFTP accounts into **AWS S3**. External tables were mounted on top of them using **AWS Athena**.
* Un-structured data such as Clickstream data was ingested into **S3** using **AWS Kinesis** and **AWS EMR** (spark structured streaming).
* Extracting and uploading data into **AWS S3** buckets using **Informatica** aws plugin.
* Used Spark Streaming **APIs** to perform transformations and actions on the fly for building common learner data model which gets the data from Kafka in near real time and persist it to **HBase**.
* Processed the real time steaming data using **Kafka**, **Flume integrating** with **Spark streaming API**.
* Consumed JSON messages using Kafka and processed the JSON file using Spark Streaming to capture UI updates.
* Experienced in working with spark ecosystem using **Spark SQL** and **Scala queries** on different formats like text file, CSV file.
* Worked with multiple storage formats (Avro, Parquet) and databases (Hive, Impala, Kudu).
* Wrote and executed various **MYSQL** database queries from Python using **Python-MySQL** connector and MySQL dB package.

**Environment:**  AWS, Apache Spark, Apache Kafka, EMR, Scala, Talend, PySpark, HBase, Hive, Sqoop, Flume, Informatica, Glue, Hadoop, Nifi, HDFS, Scala, Oozie, MySQL, Oracle 19C,IICS,Power BI.

**Cardinal Health, Dublin, Ohio Aug 2019 to Jan 2022**

**Sr. Data Engineer**

**Responsibilities:**

* Installed/Configured/Maintained Apache Hadoop clusters for application development and Hadoop tools like **Hive**, **Pig**, **Zookeeper** and **Sqoop**.
* Implemented Partitioning, Dynamic Partitions, Buckets in **HIVE**.
* Installed and Configured **Sqoop** to import and export the data into Hive from Relational databases.
* Administering large Hadoop environments build and support cluster set up, **performance tuning** and monitoring in an enterprise environment.
* Responsible for creating **Teradata** procedures to process the data from source team and then load the data to dimension/fact tables by using the ETL tool **Wherescape RED**.
* Created Teradata procedures to recalculate the email/phone preferences by using the **Wherescape RED**ETL tool.
* Performed **Oracle Golden Gate upgradations**, and cross-platform/heterogeneous environment (Operating Systems &Databases) migrations using **Oracle Golden Gate 12c/18c/19c.**
* Implemented various components in OLAP systems using different ETL tools like **OWB**, **Wherescape RED** and **Pentaho**, **Informatica PWC**.
* Provide consultation to application development teams on Oracle database technical issues and techniques.
* Close monitoring and analysis of the **MapReduce** job executions on cluster at task level and optimized Hadoop clusters components to achieve high performance.
* Developed data pipeline using **Flume**, **Sqoop**, **Pig** and **Java MapReduce** to ingest customer behavioral data into **HDFS** for analysis.
* Used **Python** & **SAS** to extract, transform & load source data from transaction systems, generated reports, insights, and key conclusions.
* Developed story telling dashboards in **Tableau** Desktop and published them on to Tableau Server which allowed end users to understand the data on the fly with the usage of quick filters for on demand needed information.
* Implemented a CI/CD pipeline with Jenkins, GitHub, Nexus, Maven and AWS AMI's.
* Build **AWS Glue** to automatic insertion of files coming from Vendor SFTP accounts into **AWS S3**. External tables were mounted on top of them using **AWS Athena**.
* Designed and Developed data mapping procedures ETL-Data Extraction, Data Analysis and Loading process for integrating data using **R programming**.
* Created different **Pig scripts** & executed them through **shell scripts.**
* Load the data into HDFS from different Data sources like **Oracle**, **DB2** using **Sqoop** and loaded into Hive tables.
* Designed and developed Pig Latin scripts and Pig command line transformations for data joins and custom processing of MapReduce outputs.
* Worked on google cloud platform (GCP) services like compute engine, cloud load balancing, cloud storage, **cloud SQL**, **stack driver** monitoring and cloud deployment manager.
* Setup Alerting and monitoring using Stack driver in **GCP**.
* Design and implement large scale distributed solutions in **AWS** and **GCP** clouds.
* Monitoring the Hadoop cluster functioning through MCS and worked on **NoSQL** databases including HBase.
* Used Hive and created Hive tables and involved in data loading and writing **Hive UDFs** and worked with Linux server admin team in administering the server hardware and operating system.
* Designed, developed and did maintenance of data integration programs in a **Hadoop** and **RDBMS** environment with both traditional and non-traditional source systems as well as RDBMS and **NoSQL** data stores for data access and analysis.
* Configured Spark Streaming to receive real time data from the **Kafka** and store the stream data to **HDFS**.

**Environment:** Hadoop YARN, Zookeeper, Spark 1.6, Spark Streaming, Spark SQL, Scala, Pig, Python, Hive, Sqoop, Map Reduce, NoSql, HBase, Tableau, Java, AWS S3, Oracle 12c,18c,19c, Linux

**Progressive Insurance, Bangalore, India Sept 2016 to Nov 2018**

**Data Engineer**

**Responsibilities:**

* Created Spark jobs by writing **RDDs in Python** and created data frames in Spark SQL to perform data analysis and stored in **Azure Data Lake**.
* Configured Spark Streaming to receive real-time data from the **Apache Kafka** and store the stream data to **HDFS using Scala.**
* Developed Spark Applications by using kafka and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* Created various data pipelines using **Spark, Scala and SparkSQL** for faster processing of data.
* Designed batch processing jobs using Apache Spark to increase speed compared to that of **MapReduce jobs.**
* Written Spark-SQL and embedded the SQL in SCALA files to generate jar files for submission onto the **Hadoop cluster**
* Developed data pipeline using **Flume** to ingest data and customer histories into HDFS for analysis.
* Executing Spark SQL operations on JSON, transforming the data into a tabular structure using data frames, and storing and writing the data to Hive and HDFS.
* Prepare Conceptual, Logical & Physical Data models covering all the entities using **Erwin Data Modeler 9.64.**
* Build **AWS Glue** to automatic insertion of files coming from Vendor SFTP accounts into **AWS S3**. External tables were mounted on top of them using **AWS Athena**.
* Expertise in Erwin Data Modeler tool for **ER**, **Dimensional Data Modelling**.
* Worked with HIVE data warehouse infrastructure-creating tables, data distribution by implementing partitioning and bucketing, writing, and optimizing the **HQL queries.**
* Created hive tables as per requirement were Internal or External tables defined with appropriate static, dynamic partitions, and bucketing, intended for efficiency.
* Experience of using **CI/CD** techniques and processes DevOps/Git repository code promotion.
* Used **Hive** as an **ETL** tool for event joins, filters, transformations, and pre-aggregations.
* Involved in moving all log files generated from various sources to **HDFS** for further processing through **Kafka**.
* Extracting real-time data using **Kafka** and **Spark streaming** by Creating DStreams and converting them into **RDD**, processing it, and stored it into.
* Used **Spark SQL** for Scala interface that automatically converts RDD case classes to schema RDD.
* Extracted source data from Sequential files, XML files, CSV files, transformed and loaded it into the target Data warehouse.
* Solid understanding of **No SQL Database (MongoDB and Cassandra).**
* Involved in converting Hive/SQL queries into Spark transformations using **Spark RDDs**, **Spark SQL** and Scala extracted large datasets from **Cassandra** and Oracle servers into HDFS and vice versa using Sqoop.
* Involved in converting Hive/SQL queries into Spark transformations using **Spark RDDs** and Scala.
* Involved in Migrating the platform from **Cloudera to EMR platform.**
* Developed analytical component using Scala, Spark and Spark Streaming.
* Worked on developing ETL processes to load data from multiple data sources to using **FLUME and** performed structural modifications using HIVE.
* Provided technical solutions on **MS AWS, Hive, HBase, MongoDB, Telerik, Power BI, Spot Fire, Tableau, SQL Data Warehouse Data Migration Techniques using BCP, Azure Data Factory, and Fraud prediction using Azure Machine Learning.**

**Environment:**Hadoop, Hive, Kafka, Snowflake, Spark, Scala, HBase, Cassandra, JSON, XML, UNIX Shell Scripting, Cloudera, MapReduce, Power BI, ETL, MySQL, NoSQL

**Way fair, Bangalore, India Jun 2014 to Aug 2016**

**Data Engineer**

**Responsibilities:**

* Involved in designing physical and logical data models using **ERwin Data modeling tool.**
* Designed the relational data model for operational data store and staging areas, Designed Dimension & Fact tables for data marts.
* Extensively used ERwin data modeler to design Logical/Physical Data Models, relational database design.
* Created **Stored Procedures, Database Triggers, Functions and Packages** to manipulate the database and to apply the business logic according to the user's specifications.
* Created **Triggers, Views, Synonyms and Roles** to maintain integrity plan and database security.
* Creation of database links to connect to the other server and Access the required info.
* Integrity constraints, database triggers and indexes were planned and created to maintain data integrity and to facilitate better performance.
* Used Advanced Querying for exchanging messages and communicating between different modules.
* System analysis and design for enhancements Testing Forms, Reports and User Interaction.

**Environment**: Oracle 9i, SQL\* Plus, PL/SQL, ERwin, TOAD, Stored Procedures.