**Name: Abilash V**

**Phone: (469) 731-4312**

**Data Engineer**

**Professional Summary**

* Around 7+ years of extensive IT experience as a data engineer with expertise in designing data-intensive applications using Hadoop Ecosystem and Big Data Analytical, Cloud Data engineering (AWS, Azure), Data Visualization, Data Warehouse, Reporting, and Data Quality solutions.
* Hands-on expertise with the Hadoop ecosystem, including strong knowledge of Big Data technologies such as HDFS, Spark, YARN, Scala, MapReduce, Apache Cassandra, HBase, Zookeeper, Hive, Oozie, Impala, Pig, and Flume.
* With the knowledge on Spark Context, Spark-SQL, Data frame API, Spark Streaming, and Pair RDD's, worked extensively on PySpark to increase the efficiency and optimization of existing Hadoop approaches.
* Good understanding of Spark Architecture including Spark Core, Spark SQL, Data Frames, Spark Streaming, Driver Node, Worker Node, Stages, Executors, and Tasks.
* In-depth understanding and experience with real-time data streaming technologies such as Kafka and Spark Streaming.
* 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.
* Proven experience deploying software development solutions for a wide range of high-end clients, including Big Data Processing, Ingestion, Analytics, and Cloud Migration from On-Premise to AWS Cloud.
* Expertise in Azure infrastructure management (Azure Web Roles, Worker Roles, SQL Azure, Azure Storage).
* Experience in Migrating SQL database to Azure data Lake, Azure Synapse, Azure data lake Analytics, Azure SQL Database, Data Bricks, and Azure SQL Data warehouse and controlling and granting database access and migrating on-premises databases to Azure Data Lake store using Azure Data factory.
* Demonstrated understanding of the Fact/Dimension data warehouse design model, including star and snowflake design methods.
* Experienced in building Snowpipe and In-depth knowledge of Data Sharing in Snowflake and Snowflake Database, Schema and Table structures.
* Designed and developed logical and physical data models that utilize concepts such as Star Schema, Snowflake Schema and Slowly Changing Dimensions.
* Expertise in using Airflow and Oozie to create, debug, schedule, and monitor ETL jobs.
* Experience of Partitions, bucketing concepts in Hive, and designed both Managed and External tables in Hive to optimize performance.
* Experience with different file formats like Avro, parquet, ORC, JSON, XML, and compressions like snappy &bzip.
* Hands-on experience in handling database issues and connections with SQL and NoSQL databases such as MongoDB, HBase, SQL server. Created Java apps to handle data in MongoDB and HBase.

**Professional Experience**

**Client: Pfizer , Manhattan, NY Apr 2020 – Present**

**AWS Data Engineer**

**Responsibilities:**

* Implemented solutions utilizing Advanced AWS Components: EMR, EC2, etc. integrated with Big Data/Hadoop Distribution Frameworks: Hadoop YARN, Map Reduce, Spark, Hive, etc.
* Used AWS Athena extensively to ingest structured data from S3 into multiple systems, including RedShift, and to generate reports.
* Created on-demand tables on S3 files using Lambda Functions and AWS Glue using Python and PySpark.
* Performed end-to-end Architecture and implementation assessment of various AWS services like Amazon EMR, Redshift, S3, Athena, Glue, and Kinesis.
* Created AWS RDS (Relational database services) to work as Hive metastore and could combine EMR cluster's metadata into a single RDS, which avoids the data loss even by terminating the EMR.
* Involved in code migration of quality monitoring tool from AWS EC2 to AWS Lambda and built logical datasets to administer quality monitoring on snowflake warehouses.
* Designed and implemented ETL pipelines on S3 parquet files on data lake using AWS Glue.
* Created monitors, alarms, notifications, and logs for Lambda functions, Glue Jobs, EC2 hosts using CloudWatch, and used AWS Glue for the data transformation, validate and data cleansing.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the fly for building the common learner data model which gets the data from Kinesis in near real-time.
* Worked with Snowflake cloud data warehouse and AWS S3 bucket for integrating data from multiple source system which include loading nested JSON formatted data into snowflake table.
* Used AWS glue catalog with crawler to get the data from S3 and perform SQL query operations and JSON schema to define table and column mapping from S3 data to Redshift.
* Worked on EMR Security Configurations, to store the self-signed certificates as well as KMS keys created into it. This makes it spin up a cluster in an easy manner without modifying permissions after the call.
* Involved in importing the real-time data using Kafka and implemented Oozie jobs for daily imports.
* Performed partitioning and Bucketing concepts in Apache Hive database, which improves the retrieval speed when someone performs a query.

**Environment**: Hadoop YARN, Snowflake, HDFS, PySpark, Spark Streaming, Kafka Spark SQL, Python, Hive, Sqoop, Tableau, AWS S3, Athena, Lambda, CloudWatch Glue, Redshift, EMR, EC2, Linux.

**Client: American Express, Norfolk, VA Aug 2018 – March 2020**

**Azure Data Engineer**

**Responsibilities:**

* Created Pipelines in ADF using Linked Services/Datasets/Pipeline to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool, and backward.
* Extracted, Transformed and Loaded data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, Spark SQL, and U-SQL Azure Data Lake Analytics.
* Data is Ingested to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Worked on Azure Services like IaaS, PaaS and worked on storage like Blob (Page and Block), SQL Azure.
* Implemented OLAP multi-dimensional functionality using Azure SQL Data Warehouse.
* The data is retrieved using Azure SQL and Azure ML is used to built, test, and predict the data.
* Worked on Cloud databases such as Azure SQL Database, SQL managed instance, SQL Elastic pool on Azure, and SQL server.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
* Involved in Migrating Objects from Teradata to Snowflake and created Snowpipe for continuous data load.
* Increased consumption of solutions including Azure SQL Databases, Azure Cosmos DB, Azure SQL.
* Created continuous integration and continuous delivery (CI/CD) pipeline on Azure that helps to automate steps in the software delivery process.
* Deploying and managing applications in Datacenter, Virtual environment, and Azure platform as well.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD's and PySpark.
* Log data Stored in HBase DB is processed and analyzed and then imported into Hive warehouse, which enabled end business analysts to write HQL queries.
* Handled importing of data from various data sources, performed transformations using Hive, and loaded data into HDFS.
* Design, development, and implementation of performant ETL pipelines using PySpark and Azure Data Factory.

**Environment**: Azure Data Factory(V2), Azure Data Bricks, Azure Data Lake, Azure BLOB Storage, Azure ML, Azure SQL, PySpark, Hive, Git, GitHub, JIRA.

**Client: John Deere,East Moline, IL. Oct 2017 – July 2018**

**Big 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 Scala and Implemented Apache Spark data processing project to handle data from various RDBMS and Streaming sources.
* 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.
* 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.
* 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 Cassandra.
* 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.
* 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.
* Developed ETL pipelines in and out of data warehouse using a combination of Python and Snowflakes SnowSQL Writing SQL queries against Snowflake.
* Worked on developing ETL processes to load data from multiple data sources to HDFS using FLUME, and performed structural modifications using HIVE.

**Environment**: Hadoop, Hive, Kafka, Snowflake, Spark, Scala, HBase, Cassandra, JSON, XML, UNIX Shell Scripting.

**Client: AllState, Charlotte, NC Sep 2016 - Oct 2017**

**Big Data Engineer**

**Responsibilities:**

* Collaborated with business user’s/product owners/developers to contribute to the analysis of functional requirements.
* Implemented Spark SQL queries that combine Hive queries with Python programmatic data manipulations supported by RDDs and data frames.
* Used Kafka Streams to Configure Spark streaming to get information and then store it in HDFS.
* Extract Real-time feed using Spark Streaming and convert it to RDD and process data in the form of Data Frame and save the data in HDFS.
* Developing Spark scripts, UDFS using Spark SQL query for data aggregation, querying, and writing data back into RDBMS through Sqoop.
* Worked on analyzing Hadoop clusters using different big data analytic tools including HBase database and Sqoop.
* Worked on importing and exporting data from Oracle, and DB2 into HDFS and HIVE using Sqoop for analysis, visualization, and generating reports.
* Creating and inserting data into Hive tables for dynamically inserting data into data tables using partitioning and bucketing for EDW tables and historical metrics.
* Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations, and others during the ingestion process itself.
* Created ETL packages with different data sources (SQL Server, Oracle, Flat files, Excel, DB2, and Teradata) and loaded the data into target tables by performing different kinds of transformations using SSIS.
* Designed, developed data integration programs in a Hadoop environment with NoSQL data store Cassandra for data access and analysis.
* Created partitions, bucketing across the state in Hive to handle structured data using Elastic search.
* Performed Sqooping for various file transfers through the HBase tables for processing of data to several NoSQL DBs- Cassandra, Mongo DB.

**Environment**: Hadoop, MapReduce, HDFS, Hive, python, Kafka, HBase, Sqoop, NoSQL, Spark 1.9, PL/SQL, Oracle, Cassandra, Mongo DB

**Client: Mega Soft Tech Feb 2013 - Nov 2015**

**Data Analyst/ Engineer**

**Responsibilities:**

* Involved in designing physical and logical data model 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.