| **SOUMITH KUMAR ANANTHULA**  **Hadoop/ Azure Data/ Cloud Data Engineer**  **Phone: 7168165140**  **Email:** [soumithkumara@gmail.com](mailto:soumithkumara@gmail.com)  **LinkedIn:**  <https://www.linkedin.com/in/soumithask-ask/>  **Certification:**  [https://www.credly.com/badge](https://www.credly.com/badges/0b79fadf-8413-4fc8-bf1c-b67c4f325d78/public_url%20) | **A blue and white logo  Description automatically generated** |
| --- | --- |

**Azure | Hadoop | Big Data Analytics | Python | Spark | Power BI | Snowflake | ETL Developer| Data Engineer| AWS| GCP| Snowflake| ETL Testing| DataBricks| ADLS**

**About Me**

Highly skilled and experienced **Senior data Engineer** with around 10 years of expertise in building data solutions using SQL Server, MSBI, and Azure Cloud technologies. Proficient in **Azure Data Factory, Data Lake Storage, Azure Synapse Analytics, Python, and Spark**. **Extensive experience in** **ETL development, data migration, and data visualization using Power BI, AWS, GCP.** Strong understanding of data warehousing concepts and data modeling. Possess a successful track record in delivering data solutions for various domains, including **E-Commerce, Financials, and Data Analytics.**

**Professional Summary:**

* Strong development knowledge in creating pipelines, data flows and complex data transformations and manipulations using **ADF** with **Databricks** and **Azure Storage Explorer.**
* I have experience in **Azure Cloud, Azure Data Factory, Databricks, Azure Data Lake Storage Azure Synapse**.
* Visualizations filters, page level filters and report level filters in **Power BI** and creating groups and Content packages and in **Power BI** and scheduling refresh using **Power BI** gateways.
* Created reports using **Power BI Desktop, Power BI mobile, Power BI Service**, and Creating **Power Bi reports & dashboards** with SQL Server/Tabular SSAS, Excel, web, csv, Azure (Azure Data Lake Store, Azure Data Lake Analytics) and other sources and **DAX Queries**.
* Worked on SQL Server **BI** suite (ETL, Reporting, Analytics, Dashboards) using SSIS, SSAS, SSRS
* Creating real-time data streaming solutions using **Apache Spark /Spark Streaming, Kafka, Flume.**
* Developing Spark application using Spark Core, Spark SQL, Spark Streaming APIs in Scala. Python.
* Developed ETL Python data pipeline jobs orchestration, workflow scheduling and monitoring with Apache Oozie, **Airflow, AWS Step Functions and AWS Lambda**.
* Migrated data from various sources to Azure, especially from AWS to Azure.
* Big Data & Hadoop Extensive experience in implementing the best practices and design patterns for the Data Lake, enterprise data warehouse, and domain-specific data marts.
* Expertise in Big Data architecture like Hadoop (MS Azure, Hortonworks, and Cloudera) distributed systems.
* Expertise in working with Big Data/ Hadoop and Yarn architecture along with various Hadoop Demons such as Job Tracker, Task Tracker, Name Node, Data Node, Resource/ Cluster Manager, and Kafka (distributed stream-processing). Experience in configuring the Zookeeper to coordinate the servers in clusters and to maintain data consistency which is important for decision-making in the process.

**Azure Hands on experience** in setting up the Azure Data factory and creating the ingestion Pipelines to pull data to Azure Data Lake Store and Azure Blob Storage. Migration of on-premises databases to Microsoft Azure environment (Blobs, Azure Data Warehouse, Azure SQL Server, PowerShell Azure components, SSIS Azure components). Skilled Azure Data Factory (ADF), Databricks, Azure Synapse, ETL, SQL Databases, Data Warehousing, PowerBI Worked on Azure Data Lakes (ADLS) ,Data Lake Analytics integrated with other Azure Services.

**GCP Collaborated** with colleagues on projects involving BigQuery, demonstrating proficiency in utilizing the platform to drive data-driven decision making. Acquired hands-on expertise in querying and manipulating large datasets efficiently using Google BigQuery's advanced features.

**TECHNICAL SKILLS:**

| **SDLC** | Agile, Scrum, Waterfall, Kanban |
| --- | --- |
| **Big Data Ecosystem** | Hadoop, MapReduce, Pig, Hive, HBase, YARN, Kafka, Flume, Sqoop, Impala, Oozie, Zookeeper, Spark, Ambari, Elastic Search, Parquet, Snappy, Airflow, NiFi |
| **Hadoop Distributions** | Cloudera (CDH3, CDH4, and CDH5), Hortonworks, MapReduce, Apache EMR |
| **Cloud Platforms** | Amazon Web Services (AWS), MS Azure |
| **MS Azure Services** | Azure SQL Database, Azure Data Lake (ADLS), Azure Data Factory (ADF), Azure SQL Data Warehouse, Azure Service Bus, Azure Synapse, Azure Key Vault, Azure Analysis Service (AAS), Azure Blob Storage, Azure Search, Azure App Service, and Azure Data Platform Services, VM’s. |
| **ETL/ BI Tools** | Informatica, SSIS, Tableau, PowerBI, SSRS |
| **ETL Tools** | Informatica Power Center 10.x/9.x/8.x (Source Analyzer, Repository Manager, Transformation Developer, Mapplet Designer, Mapping Designer, Workflow Manager, Workflow Monitor, Warehouse Designer and Informatica Server) Informatica Data Transformation B2B (Parser, Mapper & Serializer) |
| **CI/ CD** | Azure DevOps, Jenkins, Ant, Maven |
| **Ticketing Tools** | JIRA, |
| **Operating Systems** | Linux, Windows, Ubuntu, Unix |
| **Databases (RDBMS/ NoSQL)** | Oracle, SQL Server, Cassandra, Teradata, PostgreSQL, HBase, MongoDB, Milvus Vector DB. |
| **Programming Languages/Scripting** | Scala, SQL, PL/SQL, R, Python (Pandas, NumPy, SciPy, Scikit-Learn, Seaborn, Matplotlib, NLTK), Shell Scripting. |
| **DWH Schemas** | Star Schema, Snowflake Schema |
| **Data Modeling Tools** | Erwin, MS VISO |
| **Web/ Application Server** | Apache Tomcat, WebLogic, WebSphere, JBoss |
| **Version Control** | Git, Subversion, GitHub |
| **CI/CD Tools** | Jenkins, Azure DevOps |
| Reporting / BI Tools | MS Excel, Tableau, Tableau Server and Reader, Power BI, QlikView, Crystal Reports, SSRS, Splunk |
| Data Formats | CSV, Text, XML, JSON, Avro |

**PROFESSIONAL EXPERIENCE:**

| Client | ***Reckitt*,** NJ |
| --- | --- |
| Domain | digital automation |
| Project Name | data processing and analytics framework |
| Title | Sr. Azure Data Engineer |
| Duration | **June 2020 to Till Date** |

**Project Description:** Data pipelines will be built to extract, transform and load the data to Azure, which in turn will be utilized as a source data for multiple downstream systems and processes Data Wrangling, Massaging and Enrichment will be performed in the Data Lake staging area. Here we dealt with our third party data vendors **e2open** and **kinsa** who provide us with the **Marketing, Point of sales and Supply Chain.**

**Key Contributions:**

* Designed and developed a scalable data processing and analytics framework and integrated with Azure cloud services. Migrated on-prem databases and data from third party vendors which is basically marketing data and point of sales data to Azure data Lake, Azure Synapse, Azure Blob, Azure SQL Database, Azure SQL leveraging Azure Databricks and Azure Data Factory (ADF), NIFI.
* Extracted, Transformed and Loaded (ETL) data from disparate sources to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and Azure Data Lake Analytics.
* Created ETL/ELT pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Cassandra, Blob storage, Azure SQL Data warehouse, write-back to and backwards. Built data migration jobs into enterprise Azure cloud and Snowflake.
* Worked on EDL 2.0 platform where Data from multiple sources is ingested into AWS S3/ Redshift and transformed using PySpark scripts in DataBricks/ Glue.
* Used Kafka and Kafka brokers to initiate spark context and processing live streaming.
* Developed custom Kafka producer and consumer for different publishing and subscribing to Kafka topics.
* Worked on performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning. To meet specific business requirements, wrote UDF in Scala and Stored procedures.
* Designed and developed Micro Services business components and RESTful service endpoints using spring boot.
* Developed advanced ETL import packages using SSIS, T-SQL to transform and import raw index data into database tables

**Responsibilities:**

* Developed **Spark** applications using **Pyspark** and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats and loaded into the Delta Lake layer.
* Responsible for estimating the cluster size, monitoring and troubleshooting of the **Azure Databricks**
* Analyzed and built cloud data solutions using **Azure** PaaS service to support visualization of data.
* Building the pipelines to copy the data from source to destination in Azure Data Factory (ADF V2)
* Responsible for developing ETL pipelines to meet business use cases by using data flows, Azure Data Factory (ADF), Data Lake, Azure Data warehouse, Azure Synapse.
* Experience in CDL LEVEL 0 APIs-python scripting to bring data into CDL level 0, Performing validation testing with source/ destination teams.
* Developed Databricks Python notebooks to Join, filter, pre-aggregate, and process the files stored in Azure data lake storage.
* Created automated workflows with the help of triggers and Scheduled Jobs in Flows and ADF
* Manage data coming from different sources and loading of structured and unstructured data.
* Worked on different files like CSV, JSON, Flat, Parquet to load the data from source to raw tables.
* Designed and developed Power BI graphical and visualization solutions with business requirement documents and plans for creating interactive dashboards.
* Successfully creating the Linked Services on the source and as well for the destination servers
* Worked on providing row level security using powerbi to restrict data access for various users.
* Develop customized calculated Columns and measure using DAX Formula language in Power BI
* Created **Hive staging tables** and **external tables** and joined the tables as required.
* Implemented **Dynamic Partitioning**, **Static Partitioning** and **Bucketing**.
* Installed configured Hadoop Map Reduce, Hive, HDFS, Sqoop, Flume and Oozie on Hadoop cluster.
* Worked on Microsoft Azure services like **HDInsight Clusters, BLOB, ADLS, Data Factory** and Azure Logic Apps and also did POC on Azure DataBricks. Implemented Sqoop jobs for data ingestion from Oracle to Hive.

**Environment:** MS-SQL Server 2017, SQL Server Integration Services (SSIS), Azure, Azure Synapse, Power BI, Import and Export Data wizard ,Visual Studio SQL Profiler, Big Data, Milvus Vector DB, Spark, Python, NumPy, Scipy, Pandas, NLTK, Matplotlib, BeautifulSoup, PL/SQL, Cassandra and TextBlob, Scala, YARN, Spark Context, Spark-SQL, PySpark, Pair RDD's, Spark YARN, Spark MLLib, EMR, IAM, RDS, NiFi, Kafka, Oozie, Power BI, Azure, CDL data ingestion, CDL LEVEL 0 APIs-python scripting

| Client | T-Mobile |
| --- | --- |
| Domain | Telecommunication |
| Project Name | ODW/EDW Enhancements Informatica Projects, Data Migration |
| Title | ETL Developer / Data Engineer |
| Duration | **April 2018 – May 2020** |

**Project Description:**:Client is transforming todigital automation Data Analytics, Application Development, Robotic Process Automation, AI, DevOps, and Test Automation Services. Objective of the project is to create a database for a New Program and design the schemas in the data warehouse relational and dimensional. The project consists mainly of migrating data from different sources into one centralized Microsoft SQL Server Relational Database using **SSIS** while performing the **ETL** operations to support the data loads and transformations.

**Key Contributions:**

* Multi-Cloud (Azure AKS K8s, GKE K8s, Serverless) Data Processing/ML Data Processing Pipeline Architectures, SRE Site Reliability Engineering, Security Practice Implementation Preparing the end-to-end documents and architecture design on various offer and solution
* Involved in Development of PySpark applications to process and analyze text data from emails, complaints, forums, and click streams to achieve comprehensive customer care.
* Data pipelines to process streamed chunked data on AWS by ingesting from 10+ data sources.
* Automated ETL Process using PySpark & SparkSQL in EMR Clusters for reporting and data transformation. Historical Data migration of 2000 tables from on-premises (Netezza) to AWS EMR Hive and Redshift.
* Successfully led the end-to-end migration of data and workloads from Amazon Web Services (AWS) to Microsoft Azure, showcasing proficiency in cloud data migration strategies and execution.

**Responsibilities:**

* Designed and Implemented large batch jobs and streaming delta load from various source systems including traditional ERP systems. Involved in building scalable distributed data solutions using Pyspark on Cloudera Hadoop using Azure Data Factory pipelines, Azure Synapse.
* Explored PySpark framework on Azure Data Databricks and Azure Synapse for improving the performance and optimization of existing algorithms in Hadoop using PySpark Core, Spark SQL Spark Streaming APIs.
* Ingested data from relational databases to HDFS on a regular basis using Sqoop incremental import.
* Extracted structured data from relational data sources as Data Frames in Spark SQL on Databricks.
* Implemented massive transformation and scheduling on Azure DataBricks for advanced data analytics and provided data to downstream applications.
* Worked on the integration of kafka service for stream processing, website tracking, log aggregation.
* Involved in configuring and developing kafka producers, consumers, topics, brokers using java.
* 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 Kafka in near real time and Persists into Cassandra and redshift implementing massive data lake pipelines.
* Handled large datasets using Partitions, Broadcasts in pySpark, Effective & efficient Joins, Transformations and other during the ingestion process itself.
* Involved in data modeling, ingesting data into Cassandra using CQL, milvus Vector DB java APIs and other drivers.
* Involved in converting the data from Avro format to Parquet format and vice versa.
* Transformed the Data Frames as per the requirements of the data science team.
* Involved in accessing the hive tables using Hive Context and transforming the data and store it to HBase.
* Analyze the transactional data in HDFS using Hive and optimize the performance of the queries by segregating the data using clustering and partitioning.
* Experienced in working with Spark ecosystem using Spark SQL and Python queries on different formats like Text file, CSV file.
* Developed Spark Applications for various business logics using Python.
* Generated the DDLs for the tables as per Field Mapping Document in target (Snowflake) database.

**Environment:** Cloudera (CDH5), HDFS, Impala, Kudu, Parquet, Hive 2.2.0, Kafka 1.1.0, Sqoop, Shell Scripting, Spark 2.0, Azure, Glue, Redshift,ADLS, PostgreSQL, Cassandra, Pl/SQL, Linux - Centos, Map Reduce, Scala , CDL, , CDL LEVEL 0 APIs-python scripting

| Client | ***Wolters Kluwer, Indianapolis, IN*** |
| --- | --- |
| Domain | E-Trade health care financial services |
| Project Name | **MSBI Azure Development insurance** |
| Title | **Data Engineer** |
| Duration | **July 2016 -March - 2018** |

**Project Description:** E-Trade is a leading force in commercial insurance, excelling in healthcare and finance. With a history of pioneering electronic trading for individual investors, we're driving the industry's digital evolution. Our focus is on user-friendly solutions for traders, investors, and stock plans, enhancing their financial independence. In the commercial insurance realm, we empower traders and investors through a robust digital platform and expert guidance. Our flexible workflow and rules engine give lenders control over lead distribution, streamlining operations and boosting success in the sector.

**Key Contributions:**

* Developed Azure CXP Tools portal which is one solution to all Azure Customer experience teams’ problems.
* **Used SQL Azure** extensively for database needs in Customer Lookup
* Designed, developed and tested various Power BI and Tableau visualizations for dashboard and ad-hoc reporting solutions by connecting from different data sources and databases.
* Created notifications and alerts, subscriptions to reports in the Power BI service.
* Monitor, Triage, Track and resolve issues with the **Excel Power BI reports**.
* Developed **Power BI reports** using Power Query from **SharePoint** & **different Data sources**.
* using **SSIS** tools like Import and Export Wizard, Package Installation, and SSIS Package Designer.
* Prepared the complete **data mapping** for all the migrated jobs using SSIS.
* Involved in designing, developing and deploying reports in MS SQL Server environment using SSRSandSSIS in Business Intelligence Development Studio (BIDS).
* Implemented **Corp, 2FA and Role based authentication mechanism** in Azure CXP Tools which uses **Microsoft Azure Active Directory and DSTS** (Datacenter Security Token Service).

**Responsibilities:**

* Work with business process managers and be a subject matter expert for transforming vast amounts of data and creating business intelligence reports using the state-of-the-art big data technologies (**Hive, Spark, DataBricks Scoop,** and **NIFI** for ingestion of big data, **python/bash scripting /Apache Airflow** for scheduling jobs in GCP/Google’s cloud-based environments).
* Migrate Data between big query and **Azure Data Warehouse** using **ADF**, ADLS Azure Synapse and create **Cubes** on AAS with lots of complex **DAX** language for memory optimization for reporting.
* Built reports **for monitoring data loads** into GCP and driving reliability at the site level.
* Create, schedule, and manage your data integration at scale with **Azure Data Factory,** a hybrid data integration (ETL) service. Work with data wherever it lives, in the cloud or on-premises, with enterprise-grade security.
* Schedule the Data factory pipeline and per the **SLAs**. And Monitor the Tasks if anything fails need to send email to Operation and development teams and fix the issues.

***Environment:***  MS SQL Server, SQL server Reporting Services, SQL Server Integration Services (SSIS), Microsoft Visual Studio, SQL Server Management Studio, MS Excel, PL/SQL, T-SQL, Erwin, SQL Server Profiler, Database Engine Tuning Advisor, snowflake, Azure, ADF, ADLS, Azure DataBricks.

| Client | ***Maren Solutions Pvt Ltd, India*** |
| --- | --- |
| Domain | E-Commerce |
| Project Name | **Integrated Information System** |
| Title | **Big Data Engineer** |
| Duration | **Dec 2013 – Dec 2014** |

**Description:** Data Integration and Analytics project, this project is to acquire the data from multiple source systems like file feed, Oracle DB and SAP ERP. Data pipelines will be built to extract, transform and load the data to Hadoop, Which in turn will be utilized as a source data for multiple downstream systems and processes Data Wrangling, Massaging and Enrichment will be performed in the Data lake staging area.

**Key Contributions:**

* Import the data from different sources like HDFS/HBase into Spark RDD and perform computations using PySpark to generate the output response.
* Involved in creating composer instances, Dataproc clusters for cloud infrastructure and acted as site reliability engineer in bringing up environments for our workflows that run in GCP. Designed Data Marts by following Star Schema and Snowflake Schema Methodology, using industry leading Data Modeling tools.
* Snowflake database on queries and writing Stored Procedures for normalization.
* Worked with Snowflake’s stored procedures, used procedures with corresponding DDL statements, used JavaScript API to easily wrap and execute numerous SQL queries.
* Designed and Developed ETL jobs to extract data from different sources and load it in data mart in **Snowflake** and managed Snowflake clusters such as launching the cluster by specifying the nodes and performing the data analysis queries.

**Roles and Responsibilities:**

* Performed data **wrangling** and developed data **tables** for excess inventory, stock outs, future high orders, forecast inaccuracy and delayed purchase orders on supply chain data. Performed extensive **analyses** on club-item and store level attributes to determine impacts on club-item replenishment.
* Developed and automated processes for predictive model validation, deployment, and productization of **machine learning models** for stock outs prediction.
* Good Experience in Tableau Analytics. Build Dashboards for clients using Tableau.
* Experience in end to end design and deploy rich Graphic visualizations with Drill Down and Drop down menu option and Parameterized using Tableau.
* Structured the complex receivables for the customers by reconciling developing Mobile Collection, Automated Direct Debit and integrating the reports with Online Systems and Clients **ERP**s.
* Developed various solution driven views and dashboards by developing different chart types including **Pie Charts, Bar Charts, Tree Maps, Circle Views, Line Charts, Area Charts, and Scatter Plots** in **Power BI**. Created dashboard for **Transportation system financial information**which includes **revenue** and **expenditure** by routes.
* Used container systems such as Docker and container orchestration tools such as Container Service, Kubernetes, and Terraform.

**Environment:**

MS-SQL Server ,SQL Server Integration Services (SSIS), Import and Export Data wizard, TFS, SQL Server Reporting Services, Power BI, SQL Server Analysis Services (SSAS),SQL Profiler, , Python 3.0, SSIS, Azure SQL, Azure Data Lake, Azure Blob Storage, Spark

| Client | ***Oasis Infotech, India*** |
| --- | --- |
| Domain | E-Commerce, FinTech, Insurance |
| Project Name | *Support and development of ETL Programs* Compliance Management System |
| Title | *Big Data Engineer* |
| Duration | **Jun 2012 - Dec 2013** |

The objective of this project is to provide enhanced reporting for the CMS (Compliance Management System) application, increasing the effectiveness of those who need to report on managing compliance. The principle function of the reporting module was to provide both standard reporting capabilities without impacting the production database. Also, users would be able to create and run their own ad-hoc reports by selecting the “data elements”, or “objects” from within the Business Objects universe, as well as applying their own desired filters, using those same objects. Involved in production support and BO Administration activities

Data pipelines will be built to extract, transform and load the data to Hadoop, Which in turn will be utilized as a source data for multiple downstream systems and processes Data Wrangling, Massaging and Enrichment will be performed in the Data lake staging area.

**Key Contributions:**

* Created action filters, parameters, and calculated sets for preparing dashboards and worksheets using PowerBI.
* Developed Snowflake views to load and unload data from and to an AWS S3 bucket, as well as transferring the code to production.
* Developed visualizations and dashboards using PowerBI.
* Performing ETL testing activities like running the Jobs, Extracting the data using necessary queries from database transform, and upload into the Data warehouse servers.
* Created dashboards for analyzing POS data using Power BI.
* Converting Hive/SQL queries into Spark transformations using Spark data frames, Scala Python.

**Roles and Responsibilities:**

* Running Spark SQL operations on JSON, converting the data into a tabular structure with data frames, and storing and writing the data to Hive and HDFS.
* Developing shell scripts for data ingestion and validation with different parameters, as well as writing custom shell scripts to invoke spark Employment.
* Tuned performance of Informatica mappings and sessions for improving the process and making it efficient after eliminating bottlenecks.
* Worked on complex SQL Queries, cassandra procedures and convert them to ETL tasks
* Worked with PowerShell and UNIX scripts for file transfer, emailing and other file related tasks.
* Created a risk-based machine learning model (logistic regress, random forest, SVM, etc.) to predict which customers are more likely to be delinquent based on historical performance data and rank order them.

| **Education** |
| --- |

* Masters – State University of New York at Buffalo May-2016
* Major – Data Science
* Bachelors – CVR College of Engineering May- 2012
* Major- Bachelor’s in information Technology
* Certified in Web Component Development using Java EE by NIIT
* Oracle Database SQL Certified Expert
* Microsoft Certified: Azure Data Engineer Associate.