**Pavan Kumar Puvvada**

**Email:** **papuvvada@gmail.com**

**Phone: 469-769-9017**

**Cloud Data Engineer**

#### Summary:

* Data Engineer with 10+ years of total experience, including 6 years of experience in Big Data technologies.
* Experience in Cloud Computing on Azure and GCP, Managed Big Data Ecosystem Services and Data Warehousing services, API Services. Experience on tools like Azure Data Bricks, Azure Data Factory, Synapse Analytics, HDInsight, Dataflow, Data proc, Pub/Sub.
* Experience in Developing Spark applications using Spark - SQL, PySpark and Delta Lake in Databricks for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Well versed with Big data on AWS cloud services i.e. EC2, S3, Glue, Anthena, DynamoDB and RedShift.
* Worked on Hadoop, Hive, Java, Python, Scala Struts web framework.
* Hands on experience in Architecting Legacy Data Migration projects on premises to AWS cloud.
* Extensive experience in AWS cloud services such as EC2, VPC, S#, IAM, AWS SNS, AWS SQS, AWS SES, AWS SWF & AWS Direct Connect.
* Ability to perform data manipulation using python data types such as lists, tuples, sets, iterators, and generators, and Panda’s data frames.
* Experience in creating and executing Data pipelines in GCP and AWS platforms.
* Managed configuration of web app and deployed to AWS cloud server through Chef.
* Experience in building and optimizing AWS data pipelines, architectures, and data sets.
* Experience in building efficient pipelines for moving data between GCP and Azure using Azure Data factory.
* Worked on azure data factory to extract the various source data (like SAP, BLOB, Oracle etc..) by using the activities like Copy Data, Lookup, Get metadata, Stored Procedure & Foreach Loop activities.
* Experience in extracting the On-Prem database data to azure cloud environment database dynamically by using azure Data factory pipelines.
* Experience with Storm and Kafka for the real time processing of data.
* Site Reliability Engineering responbilities for Kafka platform that scales 2GB/sec and 20 Million messages/sec.
* Developed Spark code using Python/Scala and Spark- SQL for faster testing and processing of data.
* Knowledge of High Availability and Disaster Recovery options in AWS.
* Have experience in creating triggers and scheduling the ADF pipelines based on the client’s time zone.
* Have experience in creating the mount point which is used in Azure Databricks to read the ADLS Gen-2 files and for JDBC connection.
* Created instances in AWS as well as worked on migration to AWS from data center.
* Worked on magic commands in data bricks to read the various files (like Parquet, CSV, ORC, JSON) and extracted the data using the Data frames.
* Developed AWS Cloud formation templated and set up auto scaling for EC2 instances.
* Created delta tables which are used to perform the aggregations and to perform ACID transformations for the DWH.
* Experienced in working with various python integrated development environments like PyCharm, Spyder, PyStudio, PyDev, and Sublime.
* Responsible for distributed applications across hybrid AWS and physical data centers.
* Worked on configuring the Parameters in the data bricks notebook note and called the parameters in the azure data factory pipeline.
* Have experience in creating the partitions in the azure data bricks notebook to increase the performance of the data processing.
* Worked on RDDs, Data Frames, and spark streaming datasets to process the various data and loaded the data into enterprise DWH (data warehouse).
* Experience in importing and exporting data using Sqoop from HDFS to Relational Database Systems and vice versa.
* Hands on experience in Data Analytics services such as Athena, Glue, Data Catalog and Quick Sight.
* Strong expertise in the development of web-based applications using Python, Django, HTML, XML, Angular JS, CSS, REST APIs, JavaScript, JSON, and jQuery.
* Have a good understanding of the SQL table relationships and data modeling like Star schema & Snowflake Schema.
* Worked on azure Synapse Analytics to process the data warehouse data and applied the polybase method to load the Dimensions/Fact tables.
* Experience in writing HiveQL queries to store processed data into Hive tables for analysis.
* Experience in writing shell scripts to trigger Spark Jobs and experience in scheduling them as cron jobs.
* Experience in creating Apache Airflow schedulers and NIFI Flows.
* Experienced in Data Analysis, Data Profiling, Data Integration, Migration, Data governance and Metadata.
* Experience in Extraction, Transformation and Loading (ETL) data from various sources into Data Warehouses, as well as data processing like collecting, aggregating, and moving data from various sources using ADF, ADB, Power BI and Microsoft SSIS.
* Strong hands on experience with Microservices like Spring IO, Spring Boot in deploying on various cloud infrastructure like AWS.
* Write, maintain, and improve automation scripts in python and BASH.
* Design & implement migration strategies for traditional systems on Azure (Lift and shift/Azure Migrate, other third-party tools).
* Extensive use of cloud shell SDK in GCP to configure/deploy the services Data Proc, Storage, and Big Query.
* Knowledge on Management, Master Data Management and Configuration Management.
* Experience in building Pig scripts to extract, transform and load data into HDFS for processing.
* Collaborate with application architects on infrastructure as a service (IaaS) application to Platform as a Service (PaaS).
* Experience in working with Python ORM Libraries including Django ORM.
* Good knowledge on Performance tuning, Data warehousing Concepts, Data Modeling Concepts.
* Experience in moving code from development to test and test to production in the Cloud environment by using automated processes CI/CD pipelines setup.

####

#### Technical Summary:

|  |  |
| --- | --- |
| Big Data Tools on Cloud | Azure Data Factory, Data Bricks, Synapse, Analytics, Azure HDInsight, Google Big query, Dataflow, Data proc, Pub/Sub |
| Hadoop/Bigdata Technologies | HDFS, MapReduce, Hive, Pig, Sqoop, HBase, Kafka, Oozie, Flume, Nifi, CtrlM, Airflow, Zeppelin, Spark |
| Data Pipelines  | Apache Airflow, AWS Glue |
| Programming Languages | Scala, Python, SQL. |
| Cloud Services  | GCP, Azure |
| Operating Systems | Windows, Linux |
| Databases | MySQL, DB2, Oracle, Teradata, SQL Server |
| Java IDE | Eclipse |

**Projects:**

#### Client: Northern Trust Corporation Jun 2021 – Current

**Role: Cloud Data Engineer**

**Description:**

Northern Trust Corporation is a financial services company headquartered in Chicago, that caters to corporations, institutional investors, and ultra-high net worth individuals. Northern Trust has its operations in the United States and Canada. It provides personal trust, investment management, custodian banking. We were responsible for creating a robust solution using Microsoft Azure to manage the Banking and Insurance details of each customer and designed a robust data model to extract the data load into the SQL database tables. Then, using a backed data model, we created a Power BI dashboard which shows the client daily, weekly, Monthly & Yearly transactions for end user analysis purposes.

**Responsibilities:**

* Analyzing and understanding the Business requirement, creating the steps to complete the Business deliveries.
* Developing the ADF Pipelines through different sources (Blob, SQL Server, Azure Data Lake)
* Maintaining data storage in Azure Data Lake to store all the input feed files.
* Was responsible for creating on demand tables on S3 files using Lambda functions and AWS Glue using Python and PySpark.
* Written Kafka API to collect events from front end.
* Designed and implemented by configuring topics in new kafka cluster in all environment.
* Worked on Dataflows to extract the multiple source input data (like Oracle, SQL tables) and loaded into target by using join activities and derived columns and aggregations activities.
* Designed front end and backend of the application utilizing python on Django Web Framework.
* Implemented installation and configuration of multiple node cluster on cloud using AWS on EC2.
* Worked on various activities like Get metadata, Foreach Loop, Copy Data, stored procedure.
* Develop views and templated with python and Django view controller and templating language to create a user friendly website interface.
* Prototype done with HDP Kafka and Storm for click stream application.
* Created monitors, alarms, notifications and logs for Lambda functions, Glue jobs, EC2 hosts using Cloudwatch.
* Created an exception handling mechanism to all the pipelines and stored them in the error table and error file.
* Worked on dataflow activity to load the various source files like (Csv, Txt, Json).
* Worked on derived column activity to add the dynamic columns.
* Handled AWS management tools such as cloud watch and cloud trail.
* Used AWS Glue for the data transformation, validate and data cleansing.
* Successfully secured the kafka cluster with Kerberos.
* Worked on look up activity to hold the values in a table and process database by using stored procedure activity.
* Created the notebook which extracts the source file type (.csv to Parquet) format.
* Used the parquet file as input to the ingestion layer and loaded the data into staging tables.
* Used the stage tables as input to the DWH tables and applied the transformation like Joins and Merge to perform the aggregations and loaded them into the DWH tables.
* Used Python Boto 3 to configure the services AWS Glue, EC2, S3.
* Modify the existing Python/Django modules to deliver a certain format of data.
* Implemented Kafka security features using SSL and without Kerberos.
* Automated Regular AWS tasks like snapshots creation using python scripts.
* Created the input SQL config table to store all the input file names and source & target table names.
* Using the Config table as input to pass the parameters and read the values to load the data into Notebook dynamically.
* Install and configure Apache Airflow for AWS S3 bucket and create Dag’s to run the Airflow.
* Created the ADF pipeline and configured the notebook, Scheduled the pipeline as per the client time zone.
* Created parameters and variables to hold the Folder path, Filename, Values in the ADF Pipeline.
* Created the exception audit tables, which are used to capture the ADF pipeline status and source & target record count. Designed it to send an email alert, in case of any error, to the support team to look into the issue. Worked on troubleshooting the ADF pipelines on production environment to check for any high priority issues.
* Configured the azure key vault to hold the liked service secrets like database connection & Azure blob storage account access keys and SAS tokens.
* Use Python libraries NumPy and Pandas to organize and count the data, adjust the data format, and then use matplotlib to generate graphical reports.
* Installed Kerberos secured Kafka cluster with no encryption on Dev and Prod.
* Experience in setting up the whole app stack, setup, and debug log stash to send Apache logs to AWS.
* Called the WEB API data by using the web activity and extracted the API data and loaded it into SQL tables.
* Development and maintenance of data pipeline on Azure Analytics platform using Azure synapse analytics.
* Developed the Power BI dashboards which show the Sales on daily, weekly, monthly, and yearly basis and scheduled the dashboard on client’s time zone.
* Developed the stored procedure which extracted the data and performed the aggregations and loaded the data into the target table in the incremental manner.
* Created the functions which are used to extract the GL code and pass the GL value to the stored procedure.

**Environment:** SQL Server 2017, Data Warehouse, AWS, Azure Data Factory (Pipelines, Blob Storage, Configurations), Data Bricks, PySpark, Azure Data Lake, Notebooks

**Client: United Health Mar 2019 – May 2021**

**Role: Data Engineer**

**Description:**

The Honoring Customer Permissions (HCP) program is tasked with ensuring that business groups which capture customer contact information including general promotional communication permissions share that information across the enterprise so that privacy policy and laws can be appropriately honored. Microsoft Account is a system supporting several business groups of this type and which has partially integrated with the HCP systems that facilitate sharing of customer contact/permissions data across Microsoft Enterprise. However, to ensure consistency of customer contact/permissions data across Microsoft additional work is required to fully integrate Microsoft Account and HCP systems. We were responsible for getting the Insurance and policy details of claims from all the network hospitals across worldwide and the day-to-day transactions. Claims are high for the US and UK regions. So we created the ETL process which takes the data from all the network hospitals and places them in the ADLS storage. After that we extracted the data by using ADF, ADB tools to extract the data and load it into the Azure SQL database.

**Responsibilities**:

* Developed the ADF pipelines and data bricks notebook as per the business requirements.
* Worked on various activities like Get metadata, Foreach Loop, Copy Data, stored procedure.
* Developing the ADF Pipelines using different sources like CSV, Parquet, and Database.
* Written Kafka API to collect events from front end.
* Created the raw layer which holds the input feed files to store into ADLS storage account.
* Worked with AWS using EC2 for computing and S3 as storage mechanism.
* Developed a python package to connect to Teradata from Spark and generating QC Reports.
* Written Storm topology to accept the events from Kafka producer and emit into Cassandra DB.
* Created the ADF pipeline which extracted the input feed (.CSV) files and loaded them into stage SQL tables.
* By using the stage input tables, loaded the foundation layer tables by performing the ETL transformations and loaded into target tables.
* Analyzed existing database structure to determine the scope of migrations and application designs with Python frameworks.
* Used AWS Glue catalog with crawler to get the data from S3 and perform sql query operations.
* Deployed Lambda and other dependencies into AWS to automate EMR Spin for Data Lake jobs.
* Scheduled Spark applications/steps in AWS EMR cluster.
* Created the ETL logic by using activities like Look up, Foreach Loop, copy data, stored procedure activities.
* Loaded data from foundation to data mart using the data bricks notebook, applied various transformations commands.
* Created the input parameters to pass and filter the specific data set from the data source system.
* Created the mount point for ADLS Gen-2 & SQL server storage account connections.
* Created the azure key vault to store the access keys and SAS tokens to access the SQL server connection and ADLS storage connections.
* Developed a web application using Python and Django.
* Created the database configuration table to input to the data bricks notebook and extracted the list of input tables.
* Worked on Dataflow and extracted the multiple source data by using join activity, Lookup, Get metadata, derived column activity.
* Installed Ranger in all environments for second level of security in Kakfa broker.
* Have good experience on creating stored procedures and functions.
* Strong database skills in SQL to understand/debug/modify stored procedures, functions.
* Developing the Databricks Notebooks using PySpark and extracted the multiple source files and loaded them into SQL tables.
* Created Database on InfluxDB also worked on interface, created for Kafka also checked the measurements on databases.
* Created the data frames which read the input files (parquet and .csv) files and extracted the data and performed the aggregations and transformations like with column, with column rename, merge.
* Created the Delta tables by using the input files form ADLS storage account, Joined the multiple input delta tables and performed the transformations and loaded them into data mart tables.
* Created the ADF pipeline which used to schedule the data bricks notebooks based on client specified time zone.
* Created the power BI report by using column charts and bar, map & Filter which shows the daily and monthly data.
* Worked on Git CMD, Git Extensions and Visual Studio Code.

**Environment:** SQL Server 2012, 2014, AWS, Azure Data Factory

**Client: Sun Life Financial, Inc. Oct 2017 – Mar 2019**

**Role: Hadoop & Spark Developer**

**Description:**

Sun Life Financial, Inc. is a Canadian financial services company known primarily as a life insurance company. It is one of the largest life insurance companies in the world, and one of the oldest with a history spanning back to 1865. We were responsible for converting all its existing SAS code to spark and designed an Enterprise Data Lake over Hadoop as per client requirements.

**Responsibilities:**

* Understanding SAS functionality like Macros, Proc, ProcSQL's, Datasets.
* Converting SAS built in functions to Spark.
* Involved in testing the code.
* Moved files into HDFS from MySQL.
* Responsible for creating Hive Tables.
* Involved in Requirement gathering and Designing of the Data Lake.
* Create Flows to store files in Horton Works Hadoop Cluster.
* Involved in capacity planning and implementation of cluster.
* Python and Spark programs on EMR clusters for data integration.
* Developed Spark Jobs to Extract Metadata from files and store them in Apache Solr and ELK Stack.
* Created Hive tables and stored processed data into it so the end user can easily access it.
* Developed Spark Jobs to Pre-processing of Data.
* Designed the Hive tables for the data and loaded the Data using spark Jobs.
* Experience in loading data from the RDBMS database into HDFS using SQOOP.
* Created Dashboards in Apache Zeppelin Using Hive Queries.

**Client: Nationwide Mutual Insurance Company, USA Nov 2014 – Sept 2017**

**Role: Systems Engineer (ETL and BI Developer)**

**Description:**

Nationwide Mutual Insurance Company & Affiliated Companies is a group of large U.S. insurance and financial services companies based in Columbus. Nationwide is one of the largest insurance and financial services companies in the world, focusing on domestic property and casualty insurance, life insurance and retirement savings, asset management, and strategic investments.

**Responsibilities:**

* Worked on Informatica Power Center tools- Designer, Repository Manager, Workflow Manager, and Workflow Monitor
* Extensively worked on Python and build the custom ingest framework.
* Created functional specification documents to outline data flow from sources to targets.
* Used various transformations like Lookup, Filter, Expression, Sequence Generator, Update Strategy, Joiner, Stored Procedure, SQL, and Union to develop robust mappings in the Informatica Designer.
* Developed mapping parameters and variables to support SQL override and created global parameter files for loading SRC, STAGING, ODS & HIST tables.
* Responsible for loading the data from various data sources from different databases (SQL Server, MySQL, Oracle).
* Load the Flat file data to staging tables using Informatica and stored procedures.
* Developed mappings to load into Sources, STAGING, ODS and HIST tables.
* Developed mapplets where there is reusability.
* Developed mappings to load the static data into Dimensions.
* Used existing ETL standards to develop these mappings.
* Worked on different tasks in Workflows like sessions, event wait, decision, e-mail, command, and scheduling of the workflow.
* Used Debugger to test the mappings and fixed the bugs.
* Responsible for resolving QA testing issues. HP Quality Centre 11 is used to track the defect status in all phases of the project.
* Responsible for Optimizing Informatica performance using session logs.
* Prepared migration document to move the mappings from development to testing and then to production repositories.
* Would act as one point of contact for Project related issues/escalation and maintain effective communication between the IT and Business teams.
* Involved in fixing the UAT issues reported by users.
* Handled changes effectively by implementing the change management process at UAT stages.
* Responsible for updating status to the client on a weekly basis.

**Company: Micro land Ltd Nov 2011 – Oct 2014**

**Client: Western Digital - Technical Support**

**Role: Tech Support Associate**

**Description**:

Western Digital Corporation is an American computer hard disk drive manufacturer and data storage company, headquartered in San Jose, California. It designs, manufactures, and sells data technology products, including storage devices, network devices, data center systems and cloud storage services.

**Responsibilities:**

Dealt with the customers of Western Digital products in providing support on all technical aspects on

* Routers, Switches
* External/Internal Drives
* Network Drives
* Enterprise Drives
* Network Media devices