Jay Patel
(470) 610-2532 | jdpatel1905@gmail.com

SUMMARY

- AWS certified Data Engineer with experience in IT with exceptional expertise in **Big Data/Hadoop ecosystem** and **Data Analytics** techniques.
- Experience in writing queries using **SQL**, experience in **data integration** and **performance training**.
- Experienced of building ETL workflows in **Azure** platform using Azure **data bricks** and **data factory.**
- Developed various **shell scripts** and **python scripts** to automate **Spark jobs** and **Hive scripts**.
- Actively involved in all phases of data science project life cycle including **Data collection**, **Data Pre-Processing**, **Exploratory Data Analysis**, **Feature Engineering**, **Feature selection** and building **Machine learning** Model pipeline.
- Expertise in developing multiple **Kafka** Producers and Consumers as per the software requirement specifications.
- Hands on Experience in using **Visualization tools** like **Tableau, Power BI.**
- Hands on experience in using **Hadoop ecosystem** components like **Hadoop, Hive, Pig, Sqoop, HBase, Cassandra, Spark, Spark Streaming, Spark SQL, Oozie, Zookeeper, Kafka, Flume, MapReduce framework, Yarn, Scala and Hue.**
- Experience in working with **GIT**, **Bitbucket** Version Control System.
- Extensive experience working in a **Test-Driven** Development and **Agile-Scrum** Development.
- Involved in daily **SCRUM meetings** to discuss the development/progress and was active in making scrum meetings more **productive**.
- Excellent Communication skills, Interpersonal skills, problem solving skills and a team player. Ability to quickly adapt new environment and technologies.
- Proficient in **Python Scripting** and worked in stats function with **NumPy, visualization using Matplotlib and Pandas** for organizing data.
- Experience in different Hadoop distributions like **Cloudera** and **Horton Works** Data Platform (**HDP**).
- In depth understanding of **Hadoop Architecture** including **YARN** and various components such as **HDFS Resource Manager**, **Node Manager**, **Name Node**, **Data Node**.
- Hands on experience in **Importing** and **exporting** data from **RDBMS** into **HDFS** and vice-versa using **Sqoop**.
- Experience in working with **Hive data warehouse tool**-creating tables, distributing data by doing **static partitioning** and **dynamic partitioning**, **bucketing**, and using **Hive optimization techniques**.
- Experience working with **Cassandra** and **NoSQL** database including **MongoDB** and **HBase**.
- Experience in **tuning** and debugging **Spark application** and using **Spark optimization techniques**.
- Experience in building **PySpark** and **Spark-Scala** applications for **interactive analysis, batch processing and stream processing.**
- Hands on experience in creating real time **data streaming** solutions using **Apache Spark Core**, **Spark SQL,** and **Data Frames**.
- **Azure** cloud experience using **Azure Data Lake (COSMOS), Azure Data Factory, Azure Machine Learning, Azure Data Bricks.**
- Extensive knowledge in **implementing**, **configuring**, and **maintaining Amazon Web Services (AWS)** like **EC2**, **S3, Redshift, Glue** and **Athena**.
- Experience in working with **Azure** cloud platform **(HDInsight, DataLake, Databricks, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH and Data Storage Explorer)**.
- Experienced in **data manipulation** using **python** and **python libraries** such as **Pandas**, **NumPy**, **SciPy** and **Scikit-Learn** for **data analysis**, **numerical computations,** and **machine learning.**

CORE COMPETENCIES

Hadoop/Big Data Technologies	**Hadoop, Map Reduce, Sqoop, Hive, Oozie, Spark, Zookeeper and Cloudera Manager, Kafka, Flume**
ETL Tools	**Informatica**
NO SQL Database	**HBase, Cassandra, Dynamo DB, Mongo DB.**
Monitoring and Reporting	**Tableau, Custom shell scripts**
Hadoop Distribution	**Horton Works, Cloudera**
Build Tools	**Maven**
Programming & Scripting	**Python, Scala, JAVA, SQL, Shell Scripting, C, C++**
Databases	**Oracle, MY SQL, Teradata**
Machine Learning & Analytics Tools	**Supervised Learning (Linear Regression, Logistic Regression, Decision Tree, Random Forest, SVM, Classification), Unsupervised Learning (Clustering, KNN, Factor Analysis, PCA), Natural Language Processing, Google Analytics Fiddler, Tableau**
Version Control	**Git, GitHub, SVN, CVS**
Operating Systems	**Linux, Unix, Mac OS-X, CentOS, Windows 10, Windows 8, Windows 7**
Cloud Computing	**AWS, Azure**
AWS Services	**Amazon EC2, Amazon S3, Amazon Simple DB, Amazon MQ, Amazon ECS, Amazon Lambdas, Amazon Sagemaker, Amazon RDS, Amazon Elastic Load Balancing, Elastic Search, Amazon SQS, AWS Identity and access management, AWS Cloud Watch, Amazon EBS and Amazon CloudFormation**

EXPERIENCE

Client : : MassMutual, Plano, TX. **Dec 2020 – Current**

Role: Senior Big Data Engineer

Responsibilities:

- Worked extensively on Hadoop Components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node, YARN, Spark and Map Reduce programming.
- Worked extensively with Sqoop for importing and exporting the data from HDFS to Relational Database systems (RDBMS) and vice-versa.
- Written several Map reduce Jobs using Pyspark, Numpy and used Jenkins for Continuous integration.
- Created HBase tables to load large sets of structured, semi-structured and unstructured data coming from UNIX, NoSQL and a variety of portfolios.
- Optimized the Hive tables using optimization techniques like partitions and bucketing to provide better performance with HiveQL queries.
- Generated various kinds of reports using Power BI and Tableau based on Client specification.
- Developed **Spark Applications** by using **Python** and Implemented Apache Spark data processing project to handle data from various **RDBMS** and **Streaming** sources.
- Worked on cloud deployments using Maven, Docker, and Jenkins.
- Experience in using Avro, Parquet, RCFile and JSON file formats, developed UDF in Hive.
- Worked on Custom Loaders and Storage Classes in PIG to work on several data formats like JSON, XML, CSV and generated Bags for processing using PIG etc.
- Worked with **Spark** for improving performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Spark MLlib, Data Frame, Pair RDD's, Spark YARN.
- Performed tuning of Spark Applications to set batch interval time and correct level of Parallelism and memory tuning.
- 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 real time and persist it to **Cassandra.**
- Developed Kafka consumer's API in python for consuming data from Kafka topics.
- Used Kafka to consume XML messages and Spark Streaming to process the XML file to capture UI updates.

- Valuable experience on practical implementation of cloud-specific technologies including IAM, Amazon Cloud Services like Elastic Compute Cloud (EC2), ElastiCache, Simple Storage Services (S3), Cloud Formation, Virtual Private Cloud (VPC), Route 53, Lambda, Glue, EMR.
- Scheduling Spark/Scala jobs using Oozie workflow in Hadoop Cluster and generated detailed design documentation for the source-to-target transformations.
- Migrated an existing on-premises application to **AWS** and used **AWS** services like **EC2** and **S3** for small data sets processing and storage.
- Loaded data into S3 buckets using AWS Lambda Functions, AWS Glue and PySpark and filtered data stored in S3 buckets using Elasticsearch and loaded data into Hive external tables. **M**aintained and operated Hadoop cluster on **AWS EMR**.
- Used AWS EMR Spark cluster and Cloud Dataflow on GCP to compare the efficiency of a POC on a developed pipeline.
- Configured Snow pipe to pull the data from S3 buckets into Snowflakes table and stored incoming data in the Snowflakes staging area.
- Created live real-time Processing and core jobs using Spark Streaming with Kafka as a data pipe-line system.
- Worked on Amazon Redshift for shifting all Data warehouses into one Data warehouse.
- Designed columnar families in Cassandra and Ingested data from RDBMS, performed data transformations, and then exported the transformed data to Cassandra as per the business requirement.

Environment: Spark, Spark-Streaming, Spark SQL, AWS EMR, S3, EC2, MapR, HDFS, Hive, PIG, Apache Kafka, Sqoop, Python, Scala, Pyspark, Shell scripting, Linux, MySQL, NoSQL, SOLR, Jenkins, Eclipse, Oracle, Git, Oozie, Tableau, Power BI, SOAP, Cassandra, and Agile Methodologies.

Client: Change Healthcare, Nashville TN **Aug 2019 – Dec 2020**
Role: Big Data Engineer
Responsibilities:

- Created data pipeline for different events in Azure Blob storage into Hive external tables. Used various Hive optimization techniques like partitioning, bucketing and Mapjoin.
- Developed python code for different tasks, dependencies, SLA watcher and time sensor for each job for work-flow management and automation using Airflow tool.
- Developed shell scripts for dynamic partitions adding to hive stage. Involved in developing ETL jobs to extract data and load it in Datalake.
- Implemented Spark using Pyspark and Spark SQL for faster testing and processing of data
- Implemented Spark using PySpark and utilizing Data frames and Spark SQL API for faster processing of data.
- Proficient in working with Azure cloud platform (HDInsight, DataLake, DataBricks, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH and Data Storage Explorer).
- Designed and deployed data pipelines using DataLake, DataBricks, and Apache Airflow.
- Verifying Json schema change of source files and verifying duplicate files in source location. Worked on creating a query parser script in python.
- Created new features based on information from million transaction records and training models using Machine-Learning techniques such as Gradient Boosting Tree and Deep Learning techniques like RNN
- Worked on Azure Data Factory to integrate data of both on-prem (MY SQL, Cassandra) and cloud (Blob storage, Azure SQL DB) and applied transformations to load back to Azure Synapse.
- Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and U-SQL Azure Data Lake Analytics . Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
- Creating complex SQL queries and scripts to extract and aggregate data to validate the accuracy of the data and Business requirement gathering and translating them into clear and concise specifications and queries.
- Evolved in Spark Scala functions for mining data to provide real time insights and reports.
- Configured spark streaming to receive real time data from the Apache Kafka and store the stream data using Scala to Azure Table.
- DataLake is used to store and do all types of processing and analytics.

- Ingested data into Azure Blob storage and processed the data using Databricks. Involved in writing Spark Scala scripts and UDF's to perform transformations on large datasets.
- Utilized Spark Streaming API to stream data from various sources. Optimized existing Scala code and Improved the cluster performance.
- Involved in using Spark DataFrames to create Various Datasets and applied business transformations and data cleansing operations using DataBricks Notebooks.
- Efficient in writing Python scripts to build ETL pipeline and Directed Acyclic Graph (DAG) workflows using Airflow, Apache NiFi.
- Developed, validated and executed machine learning algorithms including Naive Bayes, Decision trees, Regression models, SVM, XG Boost to identify different kinds of fraud and reporting tools that answer applied research and business questions for internal clients.
- Loaded data from Web servers and Teradata using Sqoop, Flume and Spark Streaming API.
- Used Flume sink to write directly to indexers deployed on cluster, allowing indexing during ingestion.
- Migrated from Oozie to Apache Airflow. Involved in developing Oozie and Airflow.
- workflows for daily incremental loads, getting data from RDBMS (MongoDB, MS SQL).
- Monitored Spark cluster using Log Analytics and Ambari Web UI. Transitioned log storage from Cassandra to Azure SQL Datawarehouse and improved the query performance.
- Involved in developing data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL. Also Worked with Cosmos DB (SQL API and Mongo API).
- Designed custom-built input adapters using Spark, Hive, and Sqoop to ingest and analyze data (Snowflake, MS SQL, MongoDB) into HDFS.
- Used Pyspark for amazing concurrency support, and Pyspark plays the key role in parallelizing processing of the large data sets.
- Developed map reduce jobs using Pyspark for compiling the program code into bytecode for the JVM for data processing.
- Proficient in utilizing data for interactive Power BI dashboards and reporting purposes based on business requirements.
- Managed resources and scheduling across the cluster using Azure Kubernetes Service. AKS can be used to create, configure and manage a cluster of Virtual machines.
- Extensively used Kubernetes which is possible to handle all the online and batch workloads required to feed, analytics and machine learning applications.
- Used Azure DevOps and VSTS (Visual Studio Team Services) for CI/CD, Active Directory for authentication and Apache Ranger for authorization.
- Experience in working with Spark applications like batch interval time, level of parallelism, memory tuning to improve the processing time and efficiency.

Environment: Azure HDInsight, Databricks, DataLake, CosmosDB, MySQL, Azure SQL, Snowflake, MongoDB, Cassandra, Teradata, Ambari, Flume, Tableau, PowerBI, Azure AD, Git, Blob Storage, Data Factory, Data Storage Explorer, Scala, Hadoop 2.x (HDFS, MapReduce, Yarn), Spark v2.0.2, PySpark, Airflow, Hive, Sqoop, HBase, Oozie.

Client: Target, Dallas TX **Dec 2018 – Aug 2019**
Role: Big Data Engineer
Responsibilities:

- Managed multiple small projects with a team of 5 members, planned milestones, scheduled project milestones, and tracked project deliverables.
- Performed network traffic and analysis expertise using data mining, Hadoop ecosystem (MapReduce, HDFS Hive) and visualization tools by considering raw packet data, network flow, and Intrusion Detection Systems (IDS).
- Wrote Shell scripts to monitor load on database and Perl scripts to format data extracted from data warehouse based on user requirements.
- Designed, developed, and delivered the jobs and transformations over the data to enrich the data and progressively elevate for consuming in the layer of the delta lake.

- Analyzed the company's expenses on software tools and came up with a strategy to reduce those expenses by 30%.
- Created map design document to transfer data from source system to data warehouse, built ETL pipeline which made analyst's job easy and reduced the patient's expense on treatment up to 40%.
- Development of Informatica Mappings, Sessions, Worklets, Workflows.
- Created chat-bot to receive complaints from the customers and give them an estimated waiting time to resolve the issue.
- Enterprise Insurance data warehouse is a conversion project of migrating existing data marts at an integrated place to get the advantage of corporate wide data warehouse. It involves rewriting/developing existing data marts and adding new subject areas to existing data marts, it helps business users a platform queries across various subject areas using single OLAP tool (Cognos).
- Created map design document to transfer data from source system to data warehouse, built ETL pipeline which made analyst's job easy and reduced the patient's expense on treatment up to 40%.
- Created data pipeline for different events in Azure Blob storage into Hive external tables. Used various Hive optimization techniques like partitioning, bucketing and Mapjoin.
- Involved in developing data ingestion pipelines on Azure HDInsight Spark cluster using Azure Data Factory and Spark SQL. Also Worked with Cosmos DB (SQL API and Mongo API).
- Designed custom-built input adapters using Spark, Hive, and Sqoop to ingest and analyze data (Snowflake, MS SQL, MongoDB) into HDFS.
- Developed automatic job flows and ran through Oozie daily and when needed which runs MapReduce jobs internally.
- Extracted Tables and exported data from Teradata through Sqoop and placed in Cassandra.

Environment: Python, R, AWS EMR, Apache Spark, Hadoop ecosystem (MapReduce, HDFS, Hive) Scala, LogRythm, Openvas, Informatica, Ubuntu.

Client : KPIT Technologies, Mumbai, India. **Aug 2017 - Dec 2018**
Role: Data Engineer
Responsibilities:

- Extensively worked with Azure cloud platform (HDInsight, Data Lake, Databricks, Blob Storage, Data Factory, Synapse, SQL, SQL DB, DWH and Data Storage Explorer).
- Extract Transform and Load data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, T-SQL, Spark SQL and U-SQL Azure Data Lake Analytics.
- Ingested data to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
- Created Pipelines in Azure Data Factory (ADF) using Linked Services, Datasets, Pipeline to extract, transform and load data from different sources like Azure SQL, Blob storage, Azure SQL DW, write-back tool and backwards.
- Created Application Interface Document for the downstream to create new interface to transfer and receive the files through Azure Data Share.
- Designed and configured Azure Cloud relational servers and databases, analyzing current and future business requirements.
- Worked on Power Shell scripts to automate the Azure cloud system creation of Resource groups, Web Applications, Azure Storage Blobs & Tables, firewall rules.
- Worked on migration of data from On-prem SQL server to Cloud databases (Azure Synapse Analytics (DW) & Azure SQL DB).
- Configured Input & Output bindings of Azure Function with Azure Cosmos DB collection to read and write data from the container whenever the function executes.
- Designed and deployed data pipelines using Data Lake, Databricks, and Apache Airflow.
- Developed Elastic pool databases and scheduled Elastic jobs to execute T-SQL procedures.
- Developed Spark applications using PySpark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.

- Ingested data in mini-batches and performs RDD transformations on those mini-batches of data by using Spark Streaming to perform streaming analytics in Databricks.
- Created and provisioned different Databricks clusters needed for batch and continuous streaming data processing and installed the required libraries for the clusters.
- Created several Databricks Spark jobs with PySpark to perform several tables to table operations.

Environment: Azure HDInsight, Databricks, DataLake, CosmosDB, MySQL, Azure SQL, Snowflake, MongoDB, Cassandra, Teradata, Ambari, Flume, Tableau, PowerBI, Azure AD, Git, Blob Storage, Data Factory, Data Storage Explorer, Scala, Hadoop 2.x (HDFS, MapReduce, Yarn), Spark v2.0.2, PySpark, Airflow, Hive, Sqoop, HBase, Oozie.

Client : SAP Labs, Mumbai, India. **Mar 2015 - Aug 2017**
Role: Data Analyst
Responsibilities:

- Expertise in quantitative analysis, data mining, and the presentation of data to see beyond the numbers and understand trends and insights.
- Experience analyzing data with the help of Python libraries including Pandas, NumPy, SciPy and Matplotlib.
- Experience working in project with machine learning, big data, data visualization, R and Python development, Unix, SQL.
- Conduct systems design, feasibility and cost studies and recommend cost-effective cloud solutions such as Amazon Web Services (AWS).
- Creating complex SQL queries and scripts to extract and aggregate data to validate the accuracy of the data and Business requirement gathering and translating them into clear and concise specifications and queries.
- Prepared high-level analysis reports with Excel and Tableau. Provides feedback on the quality of Data including identification of billing patterns and outliers.
- Experience in working with Maps, Density Maps, Tree Maps, Heat Maps Pareto charts, Bubble chart and Bullet Chart, Piecharts, Barcharts, and Line charts.
- Worked on sort & filters of tableau like Basic Sorting, basic filters, quick filters, context filters, condition filters, top filters and filter operations.
- Experience in using the Lambda functions like filter (), map () and reduce () with pandas Data Frame
- Used Pandas API for analyzing time series. Creating regression test framework for new code.
- Identify and document limitations in data quality that jeopardize the ability of internal and external data analysts and wrote standard SQL Queries to perform data validation and created excel summary reports (Pivot tables and Charts) as well as gathered analytical data to develop functional requirements using data modeling and ETL tools.
- Read date from different sources like CSV file, Excel, HTML page and SQL and performed data analysis and written to any data source like CSV file, Excel or database.
- Worked on Django REST framework and integrated new and existing API's endpoints.
- Performed data analysis using goggle API's and created visualizations such as pie charts, waterfall charts and displayed in the web application.
- Extensive knowledge in using python libraries like OS, Pickle, NumPy and SciPy.
- Involved in using Bit bucket for version control and coordinating with the team.

Environment: *Python, PyQuery, HTML5, CSS3, Apache Spark, Django, SQL, UNIX, Linux, Windows, Oracle, NoSQL, PostgreSQL, and python libraries such as PySpark, NumPy, AWS, Bit Bucket.*

EDUCATION

Georgia State University **Atlanta, GA**
Master of Science (M.S.) in Information Systems, Concentration: Big Data Analytics