**Samir Benzada**

**Hadoop data Engineer**

(860) 263-0660

SamirZada88@gmail.com

**Professional Profile**

Professional Data Engineer with 10 years of experience working in IT with more than 7 years specialized in Big Data. Experienced with using Hadoop clusters, Hadoop HDFS, Hadoop tools and Spark, Kafka, Hive. I am highly knowledgeable in data concepts and technologies including AWS pipelines, cloud services (Amazon S3, EMR, EC2). As well as using Cassandra, HIVE, No-SQL databases (like Hbase, MongoDB), SQL databases (like Oracle, SQL, PostgresSQL, My SQL server.

**Skills:**

|  |  |  |
| --- | --- | --- |
| * Data Ingestion | * Distributed Processing | * Data Migration |
| * Data Processing | * Cloud Data Pipelines | * Data Analysis |

**Career History**

**The Hartford, Hartford, CT** January 2019 – Present

**Data Engineer**

* Creating a Pipeline from a Rest API and Store the data into MySQL Database.
* Ingesting Data using Apache Flume
* Configured Flume Agent with Rest API Keys to ingest data
* Store the data into HDFS as JSON format.
* Used Apache Spark, spark-streaming to perform transformations
* Configured a JDBC driver in PySpark to connect MySQL Database
* Wrote incremental imports into Hive tables.
* Created Dataframe to store the processed results in a tabular format.
* Installed Airflow engine to run multiple spark Jobs.
* Wrote Spark SQL queries and optimized the Spark queries with Spark SQL.
* ETL to Hadoop file system (HDFS) and wrote Spark UDFs.
* Experienced in importing real-time logs to HDFS using Flume.
* Created UNIX shell scripts to automate the build process, and to perform regular jobs like file transfers.
* Hortonworks used for installation of Hortonwork Cluster and performance monitoring.
* Made incremental imports to Hive with Sqoop.
* Managed Hadoop clusters and check the status of clusters using Ambari.
* Moved Relational Database data using Sqoop into Hive Dynamic partition tables using staging tables.
* Initiated data migration from/to traditional RDBMS with Apache Sqoop.
* Developed scripts to automate the workflow processes and generate reports

**Progressive Corporation, Mayfield, OH**  May 2017 – January 2019

**AWS Cloud Data Engineer**

* Designed and implemented test environment on AWS.
* Transferred data using Informatica tool from AWS S3.
* Using AWS Redshift for storing the data on cloud.
* Used Spark SQL to perform transformations and actions on data residing in Hive.
* AWS Cloud services planning, designing and DevOps support like
* IAM user, group, roles & policy management. AWS Access Key management
* EC2 Instance creation and Auto Scaling, snapshot backup and managing template.
* Cloud formation scripting, security and resources automation.
* Implemented Spark using Scala, and utilized DataFrames and Spark SQL API for faster processing of data.
* Involved in converting HiveQL/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Used Spark DataFrame API over Cloudera platform to perform analytics on Hive data.
* Used Spark streaming to receive real time data using Kafka
* As part of Batch Modernization initiative in EC2, Analyzed existing batch ingestion developed in Oracle Data
* Integrator and developed PySpark application as ETL tool. This reduced the batch ingestion time 2x Minutes.
* Setup cloud compute engine managed and unmanaged mode and SSH key management.
* IAM access control and policy creation, service account and access management.

**McKesson Corporation, Irving, TX** December 2015 – May 2017

**Big Data Engineer**

* Analyzed and documented existing customizations as well as the current SharePoint 2013 environment.
* Used Spark-SQL and Hive Query Language (HQL) for getting customer insights, to be used for critical decision making by business users.
* Developed a task execution framework using SQL and HiveQL.
* Used Spark SQL and Data Frame API extensively to build Spark applications.
* Used Spark engine, Spark SQL for data analysis and provided to the data scientists for further analysis.
* Performed streaming data ingestion to the Spark distribution environment, using Kafka.
* Closely worked with data science team in building Spark MLlib applications to build various predictive models.
* Used Spark-Streaming APIs to perform necessary transformations and actions on the real-time data using Kafka
* Kibana setup, dashboarding and visualization configuration.
* Used Spark Streaming to divide streaming data into batches as an input to Spark engine for batch processing.
* Configured Kafka Producer in Python to Ingest Data and send the Data to Broker Topic.
* Created a Kafka Consumer in Python to see the ingestion of the data in the console.

**Epic Systems, Verona, WI** October 2013 – December 2015

**Data Engineer**

* Manipulated and analyzed complex, high volume, and high dimensional data in Hortonworks using various querying tools.
* Created a pipeline involved in loading data from API to HDFS using Flume.
* Designed batch processing jobs using Apache Spark to increase speed compared to that of MapReduce jobs.
* Creating a new Pipeline to ingest data various sources to HBase Database
* Wrote Spark applications for data validation, cleansing, transformation, and custom aggregation.
* Developed custom aggregate functions using Spark SQL and performed interactive querying.
* Connected various data centers and transferred data between them using Sqoop and various ETL tools.
* Collecting the real-time data from Kafka using Spark Streaming and perform transformations and aggregation on the fly to build the common learner data model and persists the data into Hbase.
* Extracted the data from RDBMS (Oracle, MySQL) to HDFS using Sqoop.
* Implemented Spark DataFrame and Spark SQL API extensively for processing to replace RDD usage.

**CATERPILLAR Deerfield, IL**  November 2011 – October 2013

**Data Engineer**

* Created UNIX shell scripts to automate the build process, and to perform regular jobs like file transfers.
* Wrote shell scripts to automate workflows to pull data from various databases into FS.
* Loaded data into a shared file system using terminal.
* Performed upgrades, patches and bug fixes in Hadoop in a cluster environment.
* Wrote shell scripts for automating the process of data loading
* Created test cases during two-week sprints using agile methodology
* Designed data visualization to present current impact and growth using Power BI
* Developed a Java program to clean up data streams from server logs
* Worked with a Java Messaging system to deliver logs
* Monitored multiple SQL servers to improve query performance
* Experience in configuring, installing and managing Hadoop binaries (PoC).

**Technical Skills**

**OS:** Microsoft Windows, Microsoft Windows Server, Linux, Mac

**Programming & Scripting:** Spark, Python, Java, Scala, Hive, Kafka, SQL, HTML, CSS, JavaScript

**Database Skills:** Database partitioning, database optimization, building communication channels between structured and unstructured databases.

**Data Stores:** Data Lake, Data Warehouse, SQL Database, RDBMS, NoSQL Database, Amazon Redshift, Apache Cassandra, MongoDB, SQL, MySQL, Oracle

**Data Pipelines/ETL:** Flume, Apache Storm, Apache Spark, Nifi, Apache Kafka

**Databases:** Cassandra, Hbase, Redshift, DynamoDB, MongoDB, MS Access, SQL, MySQL, Oracle, PL/SQL, Postgres SQL, RDBMS

**HADOOP DISTRIBUTIONS:** Apache Hadoop, Cloudera Hadoop, Hortonworks Hadoop

**CLOUD PLATFORMS:** Amazon AWS - EC2, SQS, S3, Elastic Cloud, Microsoft Azure, GCP

**IDE:** Jupyter Notebooks (formerly iPython Notebooks), Eclipse, IntelliJ, PyCharm

**AWS:** AWS Lambda, AWS S3, AWS RDS, AWS EMR, AWS Redshift, AWS Kinesis, AWS ELK, AWS

**Education**

**Bachelor of Science**  
**Major: General Science**  
Cavaillon, France (2001)  
  
**Master of Electrical Engineering**  
**Major: Electrical Engineering**  
Avignon, France (2005)

**Certificates**

AWS Introduction to Amazon Elastic Compute Cloud (EC2)   
AWS Introduction to AWS Identity and Access Management (IAM)   
AWS Simple Storage Service - Basics (AWS Support)   
AWS Lambda Foundations   
IBM Hadoop Certification   
IBM Scala Certification   
IBM Big Data Certification   
Full-Stack Development Certification