**[Senior Big Data Engineer](https://www.postjobfree.com/resume/ad0igx/senior-big-data-university-of-md)**

**Location:**University of Maryland, MD, 21201

**Posted:**October 20, 2023

**Contact Info:**

hasbouneer211@gmail.com

[pdf](https://www.postjobfree.com/resume-download/ad0igx?output=pdf) [docx](https://www.postjobfree.com/resume-download/ad0igx?output=docx) [txt](https://www.postjobfree.com/resume-download/ad0igx?output=txt" \o "Download Text File)[**Email to me**](https://www.postjobfree.com/contact-candidate/ad0igx/senior-big-data-university-of-md?etr=)

Top of Form

Your Email: cs@advanceqt.com [change email](https://www.postjobfree.com/change-email)

**Subject:**Response to your resume Senior Big Data Engineer

Message 

Job Description (optional) 



Bottom of Form

**Resume:**

Saber Bounehas

Big Data/ Hadoop & Cloud Engineer

Phone: (410) 946-8219; E-mail: hasbouneer211@gmail.com

Profile

•A result-oriented Professional with 12 years of experience in information technology with 9+ years of experience in Big Data development with extensive experience as a Hadoop Developer, leveraging the power of Cloud Platforms (AWS, Azure) to design and implement scalable and high-performance data processing pipelines.

•Skilled in managing data analytics, data processing, machine learning, artificial intelligence, and data-driven projects.

•Successfully implemented, set up, and worked on various Hadoop Distributions (Cloudera, Hortonworks, Amazon AWS EMR and Azure Data Proc).

•Skilled in data ingestion, extraction, and transformation using ETL processes with AWS Glue, AWS EMR, and Azure Data Bricks.

•Expertise in performance optimization for a spark in multiple platforms like Databricks, Glue, EMR, and on-premises.

•Experience in using different Hadoop ecosystem components such as HDFS, YARN, MapReduce, Spark, Sqoop, Hive, Impala, HBase, Kafka.

•Proficiency in designing scalable and efficient data architectures on Azure, leveraging services like Azure Data Lake, Azure Data Factory, Azure Data Bricks, Azure SynapseDB, and PowerBi.

•Built and managed data pipelines using Azure Data Factory and Azure Data Bricks ensuring efficient and reliable data processing and analysis workflows.

•Skilled in configuring and optimizing Azure HDInsight clusters for big data processing.

•Utilized Azure Storage to enable fast and efficient data storage and retrieval, leveraging its powerful storage capabilities.

•Proficiency in data integration with Azure services like Azure Event Hubs and Azure IoT Hub.

•Proficiency in Azure Synapse Studio for data exploration, querying, and visualization.

•Expertise in Azure Functions for serverless data processing and event-driven architecture.

•Utilized AWS EMR and AWS Lambda to process and analyze large volumes of data in the cloud S3 Storage.

•Designed and implemented a data warehousing solution using AWS Redshift, and AWS Athena to enable efficient querying and analysis of data.

•Skilled in real-time data processing with AWS Kinesis for stream processing.

•Expertise in AWS CloudWatch for monitoring and managing AWS resources.

•Implementation of data security and access controls using AWS Identity and Access Management (IAM).

•Proficient in managing data lakes on AWS, including best practices for data lake architecture.

•Utilization of AWS CloudTrail for auditing and compliance monitoring.

•Hands-on experience with AWS Glue DataBrew for data preparation and transformation.

•Expertise in data integration with AWS services like AWS DataSync.

•Proficient in AWS QuickSight for data visualization and business intelligence.

•Created workflows using AWS Step Functions to orchestrate and manage complex data processing tasks.

•Formulated comprehensive technical strategies and devised scalable CloudFormation Scripts, ensuring streamlined project execution.

•Utilized AWS Cloud formation to manage and provision AWS resources for the data pipeline.

•Demonstrated excellence in using HiveQL, SQL databases (SQL Server, SQLite, MySQL server, Oracle SQL, as well as data lakes and cloud repositories to pull data for analytics.

•Skilled in managing data lakes on GCP, including best practices for data lake architecture.

•Hands-on experience with Google Cloud Dataprep for data preparation and transformation.

•Demonstrated excellence in using SQL dialects like BigQuery SQL, as well as data lakes and cloud repositories to pull data for analytics.

•Utilized Terraform with Google Cloud to manage and provision GCP resources for the data pipeline.

•Created workflows using Google Cloud Composer to orchestrate and manage complex data processing tasks.

•Expertise in Google Cloud Pub/Sub for scalable event-driven messaging.

•Utilization of Google Cloud Audit Logging for auditing and compliance monitoring.

•Extensive experience involving the migration of on-premises data into the cloud, along with the implementation of CI/CD pipelines like Jenkins, Code Pipeline, Azure DevOps, Kubernetes, Docker, GitHub

•Used Spark to improve Python programming skills and understanding of the architecture of Spark.

•Work with various file formats (delimited text files, click stream log files, Apache log files, Avro files, JSON files, XML Files).

•Added value to Agile/Scrum processes such as Sprint Planning, Backlog, Sprint Retrospective, Requirements Gathering and provide planning and documentation for projects.

•Communicate with multiple stakeholders and project managers.

•Motivated self-starter with Good Communication, Presentation, and Problem-Solving Skills and committed to learning modern technologies.

Technical Skills

Project Methods: Agile, Scrum, DevOps, Continuous Integration, Test-Driven Development, Unit Testing, Functional Testing, Design Thinking

IDE: Eclipse, IntelliJ, PyCharm

Big Data Systems: Amazon AWS – EC2, S3, Kinesis, Azure, Google Cloud, Cloudera Hadoop, Hortonworks Hadoop, Spark, Spark Streaming, Hive, and Kafka

Databases: Cassandra, HBase, DynamoDB, MongoDB, MS Access, SQL, MySQL, Oracle, PL/SQL, SQL, RDBMS, AWS Redshift, Amazon RDS, Teradata

Programming Languages: Python, Scala, Java, Bash

ETL Data Pipelines: Apache Airflow, Hive, Sqoop, Flume, Scala, Python, Apache Kafka, Logstash

Visualization: Tableau, Power BI, Quick Sight

Programming & Scripting: Python, Java, Scala

Apache: Hive, HBase, Flink, Hadoop, HDFS, Zookeeper, HDP.

Cluster Security: Kerberos, Ranger, IAM

Cloud: Amazon Web Services (AWS), Azure, Google Cloud Platform (GCP)

Amazon: Glue, Kinesis, EMR, MSK

Scheduler Tool: Airflow, Azure Data Factory

Spark Framework: Spark API, Spark Streaming, Spark Structured Streaming, Spark SQL

CI/CD: Jenkins, GitHub, GitLab

Professional Experience

Sr. Big Data Engineer / June 2023 – Present

Transamerica Corporation / Baltimore, Maryland

•The project involves cautious planning, execution, and constant management to migrate data from on-premises to the cloud while executing CI/CD automation with Jenkins.

•Applied Hadoop system administration using Hortonworks/Ambari and Linux system administration (RHEL 7, Centos.).

•Instigated a Hadoop Cloudera distributions cluster using AWS EC2.

•Designed cloud-based architecture for scalability to handle growing data volumes.

•Incessantly monitor the performance of the cloud-based data solution and make necessary optimizations as data volumes and workloads evolve.

•Plan and execute the final migration, which may involve a cut-over from the on-premises system to the cloud-based system.

•Continuously monitor resource utilization and optimize the infrastructure for cost-efficiency.

•Developed consumer intelligence reports based on market research, data analytics, and social media.

•Perform thorough testing of the migrated data and ETL processes to ensure data accuracy and completeness.

•Configure Jenkins as the CI/CD tool for automating deployment tasks.

•Create pipelines for building, testing, and deploying data engineering code and infrastructure as code (IaC).

•Utilized AWS Redshift to store Terabytes of data on the Cloud.

•Used Spark SQL and Data Frames API to load structured and semi-structured data into Spark Clusters.

•Wrote shell scripts for log files to Hadoop cluster through automatic processes.

•Validate that the cloud-based solution meets performance and scalability requirements.

•Designed and developed data pipelines in an Azure environment using ADL Gen2, Blob Storage, ADF, Azure Databricks, Azure SQL, Azure Synapse for analytics and MS Power BI for reporting.

•Implement cost management practices to control and optimize cloud infrastructure costs.

•Leverage PySpark's capabilities for data manipulation, aggregation, and filtering to prepare data for further processing.

•Implemented AWS Fully Managed Kafka streaming to send data streams from the company APIs to Spark cluster in AWS Databricks.

•Implement data ingestion from various sources into the AWS S3 data lake using AWS Lambda functions.

•Utilize PySpark to extract and transform data from different file formats (CSV, JSON, Parquet) stored in S3.

•Joined, manipulated, and drew actionable insights from large data sources using Python and SQL.

•Worked on AWS to form and manage EC2 instances and Hadoop Clusters.

•Implement data enrichment pipelines using PySpark to combine data from Snowflake with additional details from MongoDB.

•Used Spark-SQL and Hive Query Language (HQL) for obtaining client insights.

•Develop PySpark ETL pipelines to cleanse, transform, and enrich the raw data.

•Ingested large data streams from company REST APIs into EMR cluster through AWS kinesis.

•Streamed data from AWS Fully Managed Kafka brokers using Spark Streaming and processed the data using explode transformations.

•Integrate with MongoDB to retrieve relevant information and enrich the existing data.

•Finalized the data pipeline using DynamoDB as a NoSQL storage option.

•Liaised with data engineers, DevOps teams, and cloud specialists for a smooth transition.

Sr. Data Engineer / Jan 2022 – May 2023

ConocoPhillips, Houston, Texas

•Collaborate with data scientists and analysts to develop machine learning models using Amazon Sage Maker for tasks such as fraud detection, risk assessment, and customer segmentation.

•Use AWS S3 for data collection and storage, allowing for easy access and processing of large volumes of data.

•Experienced in using AWS Step Functions for pipeline orchestration and Amazon Kinesis for event messaging.

•Containerizing Confluent Kafka application and configured subnet for communication between containers.

•Involved in data cleaning and preprocessing using AWS Glue, with the ability to write transformation scripts in Python.

•Assess the existing on-premises data infrastructure, including data sources, databases, and ETL processes.

•Proficient in real-time data analysis with Amazon Kinesis Data Analytics and Amazon Managed Streaming for Apache Kafka (Amazon MSK).

•Recognize the target cloud platform (AWS) and determine the appropriate services for data storage and processing.

•Configure security measures, access controls, and encryption to ensure data protection during the migration process.

•Skilled in transforming data using Amazon Athena for SQL processing and AWS Glue for Python processing, including cleaning, normalization, and data standardization.

•Set up a CI/CD pipeline using Jenkins to automate the deployment of ETL code and infrastructure changes.

•Implementing and monitoring solutions with AWS Lambda, S3, Amazon Redshift, Databricks, and Amazon CloudWatch for scalable and high-performance computing clusters.

•Version controls the ETL code and configurations using tools like Git.

•Implement automated testing, including unit tests, integration tests, and data validation checks, within the CI/CD pipeline.

•Designed and implemented a data lake using Amazon S3 to store and process large volumes of data.

•Applied Amazon EC2, Amazon CloudWatch, and AWS CloudFormation in AWS.

•Monitored Amazon RDS and CPU Memory using Amazon CloudWatch.

•Used Amazon Athena to realize quicker results compared to Spark throughout information analysis.

•Created automated Python scripts to convert data from different sources and to generate ETL pipelines.

•Converted SQL queries into Spark transformations using Spark RDDs, Python, and Scala.

•Produced distributed query agents to perform distributed queries against Amazon S3.

•Worked with DevOps team to deploy pipelines in higher environments using AWS Code Pipeline and AWS Code Deploy.

•Executed Hadoop/Spark jobs on Amazon EMR using programs and data stored in Amazon S3 buckets.

•Loaded data from different sources such as Amazon S3 and Amazon DynamoDB into Spark data frames and implemented in-memory data computation to generate the output response.

•Implemented usage of Amazon EMR for processing Big Data across the Hadoop Cluster of virtual servers on Amazon EC2 and Amazon S3, and Amazon Redshift for data warehousing.

•Wrote streaming applications with Apache Spark Streaming and Amazon Managed Streaming for Apache Kafka (Amazon MSK).

•Worked/Task tracking done in Jira – Kanban Jira board and Service Now for Agile methodology.

Sr. Cloud Data Architect / May 2020 – Dec 2021

Macy’s / New York City, NY

•Developed and maintained ETL pipelines using Apache Spark and Python on Google Cloud Platform (GCP) for large-scale data processing and analysis.

•Designed and implemented efficient data models and schema designs using Big Query for optimized data querying and storage.

•Outlined a scalable and efficient data architecture that integrates Snowflake, Oracle, GCP services, and other relevant components.

•Recognized data flow patterns, considering data extraction, transformation, loading, and analytics processing.

•Utilized Google Cloud Storage and Pub/Sub for data ingestion and event-driven data processing, respectively.

•Develop ETL pipelines to extract data from Oracle databases using efficient methods like change data capture (CDC) or scheduled batch processing.

•Created data models and schema designs for Snowflake data warehouses to support complex analytical queries and reporting.

•Worked with various data sources including structured, semi-structured and unstructured data to develop data integration solutions on GCP.

•Implemented real-time data processing using Spark, GCP Cloud Composer, and Google Dataflow for streaming data processing and analysis.

•Built data ingestion pipelines (Snowflake staging) using disparate sources and other data formats to enable real-time data processing and analysis.

•Integrated data pipelines with various data visualization and BI tools such as Tableau and Looker for dashboard and report generation.

•Mentored junior data engineers and provided technical guidance on best practices for ETL data pipelines, Snowflake, Snow pipes, and JSON.

•Implement infrastructure provisioning using Terraform, ensuring consistent and repeatable environments across different stages of the project.

•Optimized ETL and batch processing jobs for performance, scalability, and reliability using Spark, YARN, and GCP Data Proc.

•Managed and optimized GCP resources such as virtual machines, storage, and network for cost and performance efficiency.

•Used Google Cloud Composer to build and deploy data pipelines as DAGs using Apache Airflow.

•Built a machine learning pipeline using Apache Spark and scikit-learn to train and deploy predictive models.

•Implemented data security measures such as access control policies and data encryption to ensure data protection and compliance with regulatory requirements.

•Collaborated with cross-functional teams including data scientists and analysts to deliver data-driven solutions.

•Developed and maintained automated testing frameworks using python for data pipelines to ensure data quality and reliability.

•Worked with stakeholders to understand business requirements and translate them into technical data engineering solutions.

Big Data Engineer / Mar 2018 – Apr 2020

Allstate Corporation / Northbrook, Illinois

•Created data frames from different data sources like Existing RDDs, Structured data files, JSON Datasets, Azure SQL Database, and external databases using Azure Databricks.

•Loaded terabytes of different level raw data into Spark RDD for data computation to generate the output response and imported the data from Azure Blob Storage into Spark RDD using Azure Databricks.

•Used Hive Context which provides a superset of the functionality provided by SQL Context and Preferred to write queries using the HiveQL parser to read data from Azure HDInsight Hive tables.

•Modeled Hive partitions extensively for data separation and faster data processing and followed Hive best practices for tuning in Azure HDInsight.

•Caching of RDDs for better performance and performing actions on each RDD in Azure Databricks.

•Developed highly complex Python and Scala code, which is maintainable, easy to use, and satisfies application requirements, data processing, and analytics using inbuilt libraries in Azure Databricks.

•Successfully loaded files to Azure HDInsight Hive and Azure Blob Storage from Oracle, SQL Server using Azure Data Factory. Environment: Azure HDInsight, Azure Blob Storage, Azure Databricks, Linux, Shell Scripting, Airflow.

•Migrated legacy MapReduce jobs to PySpark jobs using Azure HDInsight.

•Written UNIX script for ETL process automation and scheduling, and to invoke jobs, error handling and reporting, to handle file operations, to do file transfer using Azure Blob Storage.

•Worked with UNIX Shell scripts for Job control and file management in Azure Linux Virtual Machines.

•Experienced in working at offshore and onshore models for development and support projects in Azure.

•Implemented data visualization solutions using Tableau and Power BI to provide insights and analytics to business stakeholders.

Big Data Engineer / Memphis, TN

Fedex Corporation / Jan 2016 – Feb 2018

•Orchestrated Airflow/workflow in hybrid cloud environment from local on-premises server to the cloud.

•Wrote Shell FTP scripts for migrating data to Amazon S3.

•Analyzed large amounts of data sets to determine optimal way to aggregate and report on them.

•Used Oozie workflow engine to manage interdependent Hadoop jobs and automate several types of Hadoop jobs such as Java map-reduce Hive, Pig, and Sqoop.

•Produced scripts for doing transformations using Scala.

•Developed and implemented Hive custom UDFs involving date functions.

•Installed, configured, and monitored Apache Airflow cluster.

•Designed and developed web app BI for performance analytics.

•Wrote Shell scripts to orchestrate execution of other scripts and move the data files within and outside of HDFS.

•Designed Python-based notebooks for automated weekly, monthly, quarterly reporting E.T.L

•Migrated various Hive UDFs and queries into Spark SQL for faster requests.

•Designed the backend database and AWS cloud infrastructure for maintaining company proprietary data.

•Installed and configured Hive, Pig, Sqoop, and Oozie on the Hadoop cluster.

•Used Sqoop to import data from Oracle to Hadoop.

•Developed Java Web Crawler to scrape market data for Internal products.

•Wrote simple SQL scripts on the final database to prepare data for visualization with Tableau.

•Developed DAG data pipeline to on-board and change management of datasets.

•Programmed Flume and HiveQL scripts to extract, transform, and load the data into database.

•Used Kafka on publish-subscribe messaging as a distributed commit log.

•Created Airflow Scheduling scripts in Python to automate data pipeline and data transfer.

•Implemented AWS Fully Managed Kafka streaming to send data streams from the company APIs to Spark cluster in AWS Databricks

•Created a benchmark between Hive and HBase for fast ingestion.

•Configured AWS Lambda for triggering parallel Cron jobs scheduler for scraping and transforming data.

•Use Cloudera Manager for installation and management of a multi-node Hadoop cluster.

•Hands on experience in Spark and Spark Streaming creating RDDs.

•Scheduled jobs using Control-M.

Hadoop Data Engineer / Jan 2014 – Dec 2015

Net Apps Inc. / Sunnyvale, CA

•Successfully executed solutions for feeding data from numerous sources and processing the Data-at-Rest utilizing Big Data technologies such as Hadoop, Map Reduce Frameworks, HBase, Hive.

•Performed data profiling and transformation on the raw data using Pig, Python, and oracle.

•Created ETL pipelines using different processors in Apache NIFI

•Experienced developing and maintaining ETL jobs.

•Imported millions of structured data from relational databases using Sqoop import to process using Spark and stored the data into HDFS in CSV format.

•Experienced with batch processing of data sources using Apache Spark.

•Created Hive External tables and loaded the data into tables and query data using HQL.

•Used Sqoop to efficiently transfer data between databases and HDFS and used Flume to stream the log data from servers.

•Developing predictive analytics using Apache Spark Scala APIs.

•Exploring Spark for improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, spark Paired RDD, Spark YARN.

•Developed Spark code using Scala and Spark-SQL for faster testing and data processing.

Software Engineer / Jan 2010 – Dec 2013

Sanmina Corporation / Sanjose, CA

•Installation, configuration, management, and deployment of underlying File transfer cluster on various Platform Servers across the globe.

•Provide Support during any failure or exceptions or Maintenance activities and troubleshooting MFT CC, MFT Internet Server, MFT Platform Server and DNI.

•Handling File transfers across network protocols such as MFT, FTP, SFTP, SCP, UNC, and FileZilla.

•Creating Virtual alias, Nodes and Profiles for Implementation and Scripting the PPA (Post Processing Action) on the PS.

•Understanding the whole architecture of the Software and Configurations and Setting up TIBCO Internet Server, Command center and MFT Platform servers.

•Automation of Triggering Mails for Failure/Success notifications and Audits and Logs checks are performed from the Tibco Command center.

•Setting up and managing high availability and Disaster recovery for MFT.

•Deployment of Server configuration using Network protocols and Implementation of MFT including PGP Encryption of files.

Education

Master’s degree in information technology, Optical Design & Engineering

Information Technologies, Mechanics, and Optics (ITMO) university, St. Petersburgh, Russia

Bachelor’s Degree

University Mohamed Khider, Biskra, Algeria

Certificates

Machine Learning – Neural Networks & Deep Learning