**Profile**:

Leverage 12+ years of experience combination of relevant Big Data/Analytics Areas, Good Understanding of underlying infrastructure for Big Data Solutions (Clustered/Distributed Computing, Storage, Data Center Networking), Expertise in Big Data technologies in Hadoop ecosystem Hive,HDFS, MapReduce, Yarn, Kafka, HBase, Sqoop,SQL, Scala azure Data Factory, Azure Data Lake Store, Azure

Databricks., AWs , GCP , Kudu, Data Bricks notebooks, S3, GCP, Apache Kafka, Amazon MSK,Splunk and Spark

Combine technical expertise with strong business acumen and operational understanding; ensure all technical strategies and activities align with corporate goals.

**Licenses & certifications**

[**AWS Certified Solutions Architect – Professional**](https://www.credly.com/badges/466074a0-71ec-4ff5-851f-a64b080e67ae/public_url) **PMP – ongoing**

**SIX Sigma Green Built**

# Career Highlights

* Around 9+ years of data processing and analysis experience, handled high volume data and supported various technology stacks.
* Proficient in Python and skilled in leveraging Azure services for building and managing data solutions.
* Well-versed in Snowflake for modern cloud data warehousing
* Experienced to partitioning and bucketing of data, designed, and managed data and created external tables in Hive to optimize performance
* Proficient in configuring and managing CI/CD tools such as Jenkins and GitLab CI for automated builds, tests, and deployments.
* Successfully reduced data storage costs by 40% by optimizing and implementing partitioning and bucketing strategies on cloud platforms such as AWS, and Azure.
* Implemented and optimized Continuous Integration (CI) and Continuous Deployment (CD) pipelines, streamlining the software development lifecycle.
* Proficient in using Python libraries and frameworks, such as NumPy, Pandas, Django, Flask, or TensorFlow, for data analysis, or machine learning.
* Knowledge of Python ecosystem tools like pip, virtual environments, and package management.
* Experience in working with Python's standard library and popular third-party libraries to enhance productivity and efficiency.
* Experience of integrating Python with databases, APIs, or other external systems for data retrieval and manipu
* Overhauled legacy ETL processes with new data pipelines, improving data availability and reducing latency from 24 hours to near real-time.
* Experience in importing and exporting data between HDFS and RDBMS using Sqoop
* Strong experience and knowledge of real time data analytics using Spark
* Carried out POCs on migrating to Spark-Streaming to process the live data.
* Utilized Spark Core, Spark Streaming and Spark SQL API for faster processing of Responsible for data extraction and data integration from different data sources into Hadoop Data Lake by creating ETL pipelines Using Spark, MapReduce, Pig, and Hive.
* Involved in converting Hive/SQL queries into Spark transformations using Spark Data frames and Scala.
* Load the data into Spark RDD and do in memory data Computation to generate the Output response.
* Used Spark for interactive queries, processing of streaming data and integration with popular NoSQL database for huge volume of data.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Scala.
* Analyzed the SQL scripts and designed the solution to implement using Spark.
* Utilized Snowflake to implement scalable and efficient data warehousing solutions, ensuring seamless data integration, high performance, and robust data security across various projects.
* Involved in converting MapReduce programs into Spark transformations using Spark RDD in Scala.
* Experienced in Apache Spark for implementing advanced procedures like text analytics and processing using the in- memory computing capabilities written in Scala.
* Experienced in implementing Kerberos authentication protocol in Hadoop for data security.
* Experienced in writing queries and sub-queries for SQL, Hive, Impala and Spark; and used different Spark modules like Spark Core, Spark RDDs, Spark Data frame and Spark SQL.
* Experienced in converting Hive queries into Spark Transformations and Actions
* Worked on data serialization formats for converting complex objects into sequence bits by using CSS, Avro, Parquet, JSON and CSV.
* Strong command over relational databases including MySQL, Oracle, MS SQL Server, and MS Access
* Data Processing Frameworks: Apache Beam, Apache Spark, Apache Flink, Hadoop
* Data Warehousing: SQL, NoSQL, MySQL, PostgreSQL, MongoDB, Redshift
* Programming Languages: Python, SQL, Scala and PySpark
* Utilizing ETL Tools such as Gathr and NIFI

Worked with project development teams following Agile methodologies that design source code branching, release life

**Professional Experience:**

Data Engineer - Team lead

Metlife April 2022 – Till Date

New York City: Remote Responsibilities:

* Designed and implemented data pipelines and **ETL** processes using **Azure Databricks** to ingest, transform, and load data from various sources.
* Designed and implemented **ETL** pipelines using **Python/Pyspark**, integrating data from various sources into

**Azure Data Warehouse**.

* Optimized data storage and retrieval processes by transitioning legacy systems to Delta tables, achieving a 30% reduction in storage costs and doubling query performance, thereby supporting more complex and larger scale analytics.
* Optimized SQL and NoSQL database interactions within Scala applications, improving data retrieval times and system throughput.
* Developed and maintained scalable, high-performance applications using Scala, ensuring robust and efficient code through functional programming principles.
* Collaborated with front-end developers and other team members to define API requirements and deliver comprehensive solutions.
* Integrate **DBT** with Azure data services like Azure SQL Data Warehouse, Azure Synapse Analytics, and Azure Blob Storage.
* Build and manage data pipelines using Azure Data Factory, Synapse, or other Azure services in conjunction with

**DBT**.

* Designed and implemented hybrid cloud solutions, utilizing DynamoDB for NoSQL data storage while integrating with Azure services for advanced analytics and processing.
* Create and manage SQL models, macros, and seeds within **DBT**.
* Integrated third-party APIs and services to extend application functionality and provide enriched data.
* Design and implement end-to-end **machine learning pipelines** from data ingestion and preprocessing to model training, evaluation, and deployment.
* Designed and deployed a **Medallion** architecture for a large-scale data warehouse project, leading to a 30% increase in data retrieval efficiency and supporting advanced analytics capabilities across the organization.
* Implemented data pipelines and integrations on Snowflake, contributing to improved data accessibility and analytics capabilities within the organization.
* Utilized **Snowflake** to implement scalable and efficient data warehousing solutions, ensuring seamless data integration, high performance, and robust data security across various projects.
* Automated complex data pipelines using Databricks and Snowflake, significantly reducing data processing time and improving overall data pipeline efficiency.
* Utilized **Snowflake's** cloud data warehousing platform for modernizing data storage and analytics.
* Designing, implementing, and managing data flows, ensuring efficient and secure data movement using **NIFI**
* Implemented robust data integration solutions, bridging the gap between diverse data sources and destinations using NiFi and **Azure Data Factory**.
* Utilized **Azure Data Factor**y to manage and monitor data pipelines in the cloud, optimizing performance and ensuring scalability.
* Optimized **Spark** jobs for performance by tuning configurations, managing memory, and leveraging advanced Spark features such as DataFrames, RDDs, and Spark SQL.
* Optimized log management processes by automating data ingestion and indexing in **Splunk**, leading to a 40% improvement in system efficiency and a reduction in manual data handling errors.
* Built and managed data lakes and storage solutions on Azure Blob Storage, Azure Data Lake Storage, and Delta Lake for scalable and cost-effective data storage.
* Integrated data engineering best practices into machine learning pipelines, ensuring data quality and feature engineering.
* Contributed to code reviews, peer mentoring, and continuous integration (CI/CD) pipelines to maintain high development standards and fast release cycles.
* Experience on spark optimization application, broadcast variable, repartition, coalesce
* Developed efficient algorithms and coding techniques to handle deeply nested JSONs, optimizing performance and reducing processing time.
* Transform operational data for data warehouse reports in Hadoop as batch transformation “Engine”
* Developed a **Linux** script
* Adept at creating and managing comprehensive technical documentation, ensuring clarity and accessibility for both technical and non-technical stakeholders.
* Writing a complex query to apply transformation and logic to the dataset
* Utilized **Agile** Scrum Methodology to manage and organize the team with regular code review sessions.

**Environment**: Azure Databricks . Azure Data Factory , Kafka , Splunk , Nifi , SAS, Scala, Pyspark , Python , Snowflake, DB2, Hive, Sqoop 1.4.6, Impala, Tableau, Talend, Bitbucket , Linux , Windows , Intilijji , Eclips , Snowflake, Azure Cosmos and MangoDB,Bit DbVisualizer, and IBM maistro , Airflow and UC4

Data Engineer

Corning Apr’2019 – March 2022

New York City Responsibilities:

* Designed and implemented data pipelines and ETL processes using Azure Databricks to ingest, transform, and load data from various sources.
* Designed and implemented ETL pipelines using Python/Pyspark, integrating data from various sources into Azure Data Warehouse.
* Implemented Spark jobs in Scala/Python to process and analyze terabytes of structured and unstructured data.
* Managed cloud-based data warehousing solutions on Azure, optimizing performance and scalability.
* Implemented monitoring solutions for NiFi and Azure Data Factory workflows, ensuring proactive identification and resolution of potential issues.
* Utilized Spark SQL and Hive to query structured data, create temporary views, and perform complex transformations and aggregations.
* Automated deployment and monitoring of Spark jobs using Airflow and Jenkins,
* Implemented data transformations within Azure Data Factory to adapt data to the desired format and structure.
* Optimized data pipelines for performance and efficiency, utilizing best practices and performance tuning techniques.
* Collaborated with cross-functional teams to understand data requirements and design solutions that align with business objectives.
* Maintained comprehensive documentation for NiFi and Azure Data Factory workflows, facilitating knowledge transfer and future maintenance.
* Implemented data flow solutions using Apache NiFi to automate the ingestion, transformation, and routing of data.
* Configured and optimized NiFi processors to ensure efficient data movement within the data pipeline.
* Builds a metadata system where all available data is maintained and cataloged
* Develop of data pipelines that translate raw data into powerful features and signals.
* implemented solutions for ingesting data from various sources and processing the Data-at-Rest utilizing Big
* Explored with the Spark improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark -SQL, Data Frame, PairRDD's, SparkYARN.
* Worked on Big Data Integration &Analytics based on Hadoop, Spark , SQL , Scala .
* Developed Sparkscripts using ScalaShell commands as per the requirement.
* Developed and designed automation framework using Scala and Shell scripting.
* Load the data into SparkRDD and do in memory data Computation to generate the Output response.
* Worked on major components in Hadoop Ecosystem including Hive, PIG, HBase, HBase-Hive Integration, Scala, Sqoop and Flume.
* Implemented data quality checks and validations within Spark jobs to ensure data accuracy and integrity.
* Developed Hive Scripts, Pig scripts, UNIX Shell scripts, programming for all ETL loading processes and converting the files into parquet in the Hadoop File System.
* Implemented Serverless functions using Azure platforms
* Successfully optimized Serverless functions for reduced cold start times, enhancing user experience by minimizing latency.
* Designed and deployed Serverless architectures for data-intensive tasks, such as data processing, ETL pipelines, and real-time event processing.
* Utilized Agile Scrum Methodology to help manage and organize a team of 4 developers with regular code review sessions.
* Utilized Agile Scrum Methodology to manage and organize the team with regular code review sessions.

**Environment**: Azure Databricks . Azure Data Factory , Nifi , , Scala, Pyspark , Python , Snowflake, DB2, Hive, snowflake, Sqoop 1.4.6, Impala, Tableau, Talend, Bitbucket , , Linux , Windows , Intilijji , Eclips , Azure Cosmos and MangoDB,Bit DbVisualizer, Tableau.

Data Engineer

3M July 2014 – Mar’2019

Saint Paul ,Minnesota Responsibilities:

* Implemented Spark using Scala and utilizing Data frames and Spark SQL API for faster processing of data
* Converted existing MapReduce jobs into Spark transformations and actions using Spark RDDs, Data frames and Spark SQL APIs.
* Worked on Big Data infrastructure for batch processing as well as real-time processing. Responsible for building scalable distributed data solutions using Hadoop.
* Developed real time data processing applications by using Scala and Python and implemented Apache Spark Streaming from various streaming sources like Kafka.
* Developed Spark jobs and Hive Jobs to summarize and transform data
* Expertise in implementing Spark Scala application using higher order functions for both batch and interactive analysis requirement.
* Experienced in developing Spark scripts for data analysis in Scala.
* Used Spark-Streaming APIs to perform necessary transformations.
* Involved in converting Hive/SQL queries into Spark transformations using Spark SQL and Scala.
* Worked with spark to consume data from Kafka and convert that to common format using Scala.
* Worked extensively with importing metadata into Hive and migrated existing tables and applications to work on Hive and Spark.
* Converted existing MapReduce jobs into Spark transformations and actions using Spark RDDs, Data frames and Spark SQL APIs.
* Wrote new spark jobs in Scala to analyze the data of the customers and sales history.
* Developed Scala based Spark applications for performing data cleansing, data aggregation, de- normalization and data preparation needed for machine learning and reporting teams to consume.
* Worked on troubleshooting spark application to make them more error tolerant.
* Involved in HDFS maintenance and loading of structured and unstructured data and imported data from mainframe dataset to HDFS using Sqoop and written the Spark Script to process the HDFS data.
* Used Spark API over Hadoop YARN to perform analytics on data in Hive.
* Extensively worked on the core and Spark SQL modules of Spark.
* Involved in Spark and Spark Streaming creating RDD's, applying operations -Transformation and Actions.
* Created partitioned tables and loaded data using both static partition and dynamic partition method.
* Implemented POC's on migrating to Spark-Streaming to process the live data.
* Executed Hive queries on Parquet tables stored in Hive to perform data analysis to meet the business requirements.
* Ingested data from RDBMS and performed data transformations, and then export the transformed data to HDFS as per the business requirement.
* Used Impala to read, write and query the data in HDFS.
* Worked on troubleshooting spark application to make them more error tolerant.
* Developed MapReduce and Spark jobs to discover trends in data usage by users.
* Implemented Spark using Python and Spark SQL for faster processing of data.
* Implemented algorithms for real time analysis in Spark
* Used Spark for interactive queries, processing of streaming data and integration with popular NoSQL
* Stored the output files for export onto HDFS and later these files are picked up by downstream systems.
* Utilized Agile Scrum Methodology to manage and organize the team with regular code review sessions.

**Environment:** Hadoop YARN, Spark 1.6, Spark Streaming, Spark SQL, PyTorch, Scala, Kafka, Python, Hive, Sqoop 1.4.6, Impala, Tableau, Talend, Oozie, Control-M, Oracle 12c, Snowflake and Linux

**Data Engineer Delta Airlines**

**Atlanta, Georgia March 2012 – June 2014**

Responsibilities

* Developed highly optimized Spark applications to perform various data cleansing, validation, transformation and summarization activities according to the requirement
* Data pipeline consists Spark, Hive and Sqoop and custom build Input Adapters to ingest, transform and analyze operational data .
* Involved in converting Hive/SQL queries into Spark transformations using Spark Data Frames and Scala .
* Used different tools for data integration with different databases and Hadoop .
* Analyzed the SQL scripts and designed the solution to implement using Scala .
* Used Spark for interactive queries, processing of streaming data and integration with popular NoSQL database for huge volume of data .
* Built real time data pipelines by developing Kafka producers and spark streaming applications for consuming .
* Ingested syslog messages parse them and streams the data to Kafka .
* Handled importing data from different data sources into HDFS using Sqoop and performing transformations using Hive, Map Reduce and then loading data into HDFS\ .
* Exported the analyzed data to the relational databases using Sqoop, to further visualize and generate reports for the BI team .
* Collecting and aggregating large amounts of log data using Flume and staging data in HDFS for further analysis
* Analyzed the data by performing Hive queries (Hive QL) to study customer behavior .
* Helped Dev ops Engineers for deploying code and debug issues .
* Used Hive to analyze the partitioned and bucketed data and compute various metrics for reporting .
* Developed Hive scripts in Hive QL to de-normalize and aggregate the data .
* Scheduled and executed workflows in Oozie to run various jobs .
* Experience in using Hadoop ecosystem and processing data using Amazon AWS. Developed simple to complex MapReduce streaming jobs using Java language for processing and validating the data .
* Developed data pipeline using MapReduce, Flume, Sqoop and Pig to ingest customer behavioral data into HDFS for analysis .
* Developed MapReduce and Spark jobs to discover trends in data usage by users .
* Implemented Spark using Python and Spark SQL for faster processing of data .
* Implemented algorithms for real time analysis in Spark
* Used Spark for interactive queries, processing of streaming data and integration with popular NoSQL database for huge volume of data .
* Used the Spark -Cassandra Connector to load data to and from Cassandra .
* Real time streaming the data using Spark with Kafka and SOA
* Handled importing data from different data sources into HDFS using Sqoop and also performing transformations using Hive, MapReduce and then loading data into HDFS .
* Exported the analyzed data to the relational databases using Sqoop, to further visualize and generate reports for the BI team
* Collecting and aggregating large amounts of log data using Flume and staging data in HDFS for further analysis
* Analyzed the data by performing Hive queries (HiveQL) and running Pig scripts (Pig Latin) to study customer behavior .
* Used Hive to analyze the partitioned and bucketed data and compute various metrics for reporting .
* Created HBase tables and column families to store the user event data .
* Written automated HBase test cases for data quality checks using HBase command line tools .
* Used Impala to read, write and query the Hadoop data in HDFS from HBase or Cassandra .
* Used Tez framework for building high performance jobs in Pig and Hive .
* Developed end to end data processing pipelines that begin with receiving data using distributed messaging systems Kafka through persistence of data into HBase
* Continuous monitoring and managing the Hadoop cluster using Cloudera Manager .
* Used JUnit framework to perform Unit testing of the application
* Developed interactive shell scripts for scheduling various data cleansing and data loading process .
* Performed data validation on the data ingested using MapReduce by building a custom model to filter all the invalid data and cleanse the data.
* Involved in Hive/SQL queries performing spark transformations using Spark RDDs and Python (pyspark).
* Created a Serverless data ingestion pipeline on AWS using lambda functions.

**Environment**: HP ALM, Selenium WebDriver, JUnit, Cucumber, Angular JS, Node.JS Jenkins, PyTorch GitHub, Windows, UNIX, Agile, MS SQL, IBM DB2, Putty, WinSCP, FTP Server, Notepad++, C#, DB Visualizer, CICS.

**Linkswave :**

Project Manager Jun.2010 to Jan. 2012

NY

Responsibilities:

* Manages the entire project lifecycle of technology projects in collaboration with stakeholders.
* Manage critical business activities supporting assigned projects
* Coordinate internal cross-functional resources and vendors for the execution of project activities and deliverables
* Track submissions including any external stakeholders
* Lead Implementation of complex projects to include but is not limited to: Project Management Information System, SharePoint, Enterprise Resource Planning system, etc.
* Coordinates and leads multiple project teams and vendors to ensure timely and cost-effective planning and execution of several inter-dependent projects.
* Cultivates collaborative relationships among project stakeholders to enable scope, schedule, and financial resource decisions.
* Ensures adherence to the Project Management Division’s policies, standards, and processes and identifies

opportunities for improvement.

* Develops project plans and schedules with well-defined milestones and present project plans to senior leadership for comment and approval.
* Manages communication of project status within the project team and external to the project team (customers & stakeholders); escalates issues as necessary.
* Assesses project quality and value through monitoring of incoming deliverables in comparison to Service Authority business requirements.
* Develops tools and processes to improve project value.
* Manages project financial and forecasting information and report performance to senior leaders.
* Identifies and manages project risks, defines opportunities for improvement, and works with the project team and senior leaders to establish corrective actions.
* Provides performance reports and other relevant analytics.

**Personal Data:**

**Nationality :** US Citizen

**Academic Qualification :** BS in Computer Science

**LinkedIn :** [linkedin.com/in/wael-g-b7b835264](https://www.linkedin.com/in/wael-g-b7b835264)

**Wael Gheith**