MELVIN BANDI

Toronto, Canada
melvinbandi@gmail.com, **PH: 647 655 5335**

Professional Summary

- Senior Data Engineer with 10+ years of overall IT experience in a variety of industries, which includes hands on experience in **Big Data** technologies
- Have 8+ years of comprehensive experience in Big Data processing using Hadoop and its ecosystem (**AWS, S3, Kinesis, Airflow, PySpark, Openshift, Kafka streaming and HBase, Snowflake, Hive, Iceberg, Hudi, Delta Lake, AWS Lake formation, Snowflake Apache NiFi, PowerBI, Starburst Trino, Python, SQL)**
- Worked on installing, configuring, and administrating Hadoop cluster for distributions like **Cloudera/Hortonworks** Distribution
- Efficient in writing **MapReduce** Programs and using Apache Hadoop API for analyzing the structured and unstructured data
- Expert in working with **Hive** data warehouse tool-creating tables, data distribution by implementing partitioning and bucketing, writing and optimizing the **HiveQL** queries
- Debugging **Pig** and Hive scripts and optimizing MapReduce job and debugging Map reduce job
- Administrator for **Pig**, **Hive** and **Hbase** installing updates patches and upgrades
- Hands-on experience in managing and reviewing **Hadoop logs**
- Good knowledge about **YARN** configuration
- Expertise in writing Hadoop Jobs for analyzing data using Hive QL (Queries), Pig Latin (Data flow language), and custom MapReduce programs in Java
- Extending Hive and Pig core functionality by writing custom UDFs
- Experience in importing and exporting data using **Sqoop** from HDFS to **Relational Database** Systems and vice-versa
- Hands on experience in configuring and working with **Flume** to load the data from multiple sources directly into HDFS
- Good working knowledge on **NoSQL** databases such as **Hbase**, **MongoDB** and **Cassandra**.
- Used **Hbase** in accordance with PIG/Hive as and when required for real time low latency queries
- Knowledge of job workflow scheduling and monitoring tools like **Oozie** (hive, pig) and **Zookeeper** (Hbase)
- Good working experience on **PySpark** (spark streaming, spark SQL), **Scala** and **Kafka**.
- Worked on reading multiple data formats on HDFS using **Scala**
- Involved in converting Hive/SQL queries into Spark transformations using **PySpark RDDs** and **Scala**
- Good experience in creating and designing data ingest pipelines using technologies such as Openshift
- Integrated Apache Storm with Kafka to perform web analytics. Uploaded click stream data from Kafka to Hdfs, Hbase and Hive by integrating with **Storm**
- Developed various **shell scripts** and **python scripts** to address various production issues.
- Developed and designed automation framework using **Python** and Shell scripting
- Generated **Java APIs** for retrieval and analysis on No-SQL database such as **HBase** and **Cassandra**
- Experience in **AWS EC2**, configuring the servers for Auto scaling and Elastic load balancing.
- Configuring AWS EC2 instances in VPC network & managing security through IAM and Monitoring servers health through Cloud Watch.
- Strong experience in working with Amazon EMR and setting up environments on Amazon AWS EC2 instances
- Worked and learned a great deal from AWS Cloud services like EC2, S3, EBS, and EMR.
- Good Knowledge of data compression formats like **Snappy, Avro**
- Hands on experience in developing the applications with PL/SQL, Oracle10g and MS-SQL Server RDBMS

Technical Skills

Big Data	Openshift, Cloudera, Hortonworks, HDFS, PIG, SQOOP, Hbase, Hive, Airflow, Oozie, Kafka, PySpark, AWS , Redshift , Python , Spark , SQL , data pipelines, Data Lake
Visualization	D3.js, ChartJS, React, plotly, seaborne, matplotlib, esquisse, ggplot2
Languages	Java, J2EE, SQL, PYTHON, Scala, R, SAS
Databases	IBM DB2, Oracle, SQL Server, MySQL, PostgreSQL, HBASE , MONGODB
Cloud Platforms	Snowflake, BigQuery, Databricks
Data Engineering	Python, dbt, Kafka, Parquet, Fivetran, Hightouch
Subscription & Authentication Models	Paywalls, Subscription-Based Products
Cloud	AWS, AZURE, Google Cloud Platform, Snowflake
ETL	Talend ETL, Talend Studio
Containerization	Docker, Kubernetes

Professional Experience

Manulife **Feb 2025 – Present**

Sr Data Engineer

Project Summary: Led and executed multiple data-driven projects, encompassing data management, engineering, and analytics to optimize business decision-making. Spearheaded the development of robust data pipelines and ETL processes, ensuring seamless data integration across various systems. Utilized data and Analytics tools such as Snowflake, AWS , Redshift , Python , Spark , SQL , data pipelines to create ETL pipelines to build reports that empowered stakeholders with actionable insights.

- Led the development of data pipelines and data engineering solutions to support real-time analytics
- Implemented data governance strategies ensuring compliance with privacy standards and best practices.
- Hands-on experience building applications, data platform and pipelines in cloud-native technologies. Deep technical understanding of Data and Analytics paradigms and technologies - Cloud (GCP, AWS, Azure), Databases/Warehouses (Snowflake, Oracle), Hadoop, etc.
- Developed SQL-based data models for accurate querying, reporting, and data transformation processes.
- In-depth knowledge of cloud native technologies, including AWS services like S2, EC2, EKS, Glue, Sage maker, Athena and Redshift.
- Coach and mentor CIHI staff in data conversion techniques.
- Ensure interoperability, performance and scalability, reliability and availability, data consistency, cost effectiveness, and technological risks are all addressed in the architecture;
- Formulate strategies for identifying and designing reusable components
- Work with project teams using R/python and support technology to ensure deliverables meet business, functional and technical requirements
- Provide oral and written briefings throughout the project to appraise the Project Manager and management team as to the status of activities, challenges and project risks.
- Participate in business requirements facilitation sessions and identifies and documents architectural significant requirements.
- Maintain Current knowledge of advancements in all areas of R, Python, Hadoop and EMR technologies

CIHI **Aug 2024 – Jan 2025**

Sr Data Engineer

Project Summary: Designed and implemented data governance strategies, focusing on data quality, security, and compliance throughout the data lifecycle. Collaborated with cross-functional teams to build scalable, efficient systems that support self-service analytics and streamline decision-making processes. Applied strong business acumen to align data solutions with organizational goals and communicated complex technical concepts to diverse stakeholders, driving impactful outcomes.

Responsibilities:

- Managed data products from concept to deployment, ensuring alignment with customer needs and business goals.
- Defined the data lifecycle, including data instrumentation, self-service interfaces, and reporting solutions.
- Worked cross-functionally with data engineering and analytics teams to design scalable and efficient data systems.
- Identify, design, and implement internal process improvements: automating manual processes, optimizing data delivery, re-designing infrastructure for greater scalability, data quality checks, minimize Cloud cost, etc.
- Build the infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources using PySpark, SQL, DataBricks, No-SQL, AWS
- Build analytics tools that utilize the AWS data pipeline to provide actionable insights into customer acquisition, operational efficiency and other key business performance metrics.
- Development experience with HL7 FHIR, HL7 CDA, and HL7 v2
- Work with other data engineers, data ingestion specialist, & subject matter experts across the company to consolidate methods and tool standards where practical
- Create and maintain optimal data pipeline architecture.
- Facilitate and lead detailed HL7 v2, HL7 CDA and FHIR interface mapping discussions with clients.
- Efficient use and understanding of data warehouse, data lake, OLAP and OLTP applications

Roche **May 2023 – Jul 2024**

Sr Data Engineer

Project Summary: The Project Involves Data Exploration using Spark SQL and Spark RDD on the data set from various sources. I am working with PySpark SQL for combining it with ETL applications, real time analysis of data, performing batch analysis, analysis and optimization of the queries, creating visualizations and processing of graphs and setting up the AWS platform.

Responsibilities:

- Utilized Google Analytics, Adobe Analytics, and Heap to collect and analyze user behavior, web traffic, and conversion metrics, providing actionable insights to improve user engagement and business performance.
- Developed custom tracking solutions and dashboards using web analytics tools to support data-driven decision-making and measure campaign effectiveness.
- Involved in creating Hive tables and written multiple Hive queries to load the hive tables for analyzing the market data coming from distinct sources
- Created extensive SQL queries for data extraction to test the data against the various databases
- Responsible for importing data to HDFS using Sqoop from different RDBMS servers and exporting data using Sqoop to the RDBMS servers after aggregations for other ETL operations

- Created Partitioning, Bucketing, Map side Join, Parallel execution for optimizing the hive queries
- Shared knowledge and expertise across the organization; Provides learning opportunities and knowledge sharing to team related to advanced analytics, machine learning methods and data visualization.
- Developed Simple to complex Map Reduce Jobs using Hive and Hbase
- Orchestrated various **Sqoop queries, Pig scripts, Hive queries** using Oozie workflows and sub-workflows
- Responsible for handling different data formats like **Avro, Parquet and ORC** formats
- Involved in generating analytics data using Map/Reduce programs written in **Python**
- Worked with Apache Airflow for Orchestrating complex computational workflow and data processing pipelines
- Involved in creation and designing of data ingest pipelines using technologies such as Apache Kafka, stream processing systems like Storm, Event Hub, IoT Hub
- Responsible for developing multiple **Kafka** Producers and Consumers from scratch as per the software requirement specifications.
- Experience in custom aggregate functions using **Spark SQL** and performed interactive querying.
- Implemented Apache **Spark** data processing project to handle data from RDBMS and streaming sources.
- Worked on **Spark SQL** and Data frames for faster execution of Hive queries using Spark
- Performed analysis on implementing Spark using **Scala**
- Designed batch processing jobs using **Apache Spark** to increase speeds by ten-fold compared to that of MR jobs
- Configuring AWS EC2 instances in VPC network & managing security through IAM and Monitoring servers health through Cloud Watch.
- Active member for developing POC on streaming data using **Apache Kafka and Spark** Streaming
- Involved in daily **SCRUM** meetings to discuss the development/progress and was active in making scrum meetings more productive

Environment: Hadoop, Kinesis, AWS, AIRFLOW, HDFS, Hive, Pig, PySpark, Impala, Scala, Kafka, Shell Scripting, Eclipse, MySQL, Talend, HBASE, Snowflake

Mobileum **Aug 2022 – Apr 2023**

Big Data Engineer

Project Summary: This project involves mainly in getting larger data sets and processing the incoming files. This application provides the implementation of Hadoop cluster and data integration for developing large scale applications, handling large datasets using Pyspark on Snowflake

Responsibilities:

- Managed and optimized cloud-based data storage and processing in **Snowflake**, **BigQuery**, and **Databricks** for large-scale data warehousing, ensuring fast and secure access to data for analytics and reporting.
- Led the migration of on-prem data solutions to cloud-based platforms, improving data accessibility and system performance across teams.
- Configured deployed and maintained multi-node Dev and Test Kafka Clusters.
- Developed Spark scripts by using Java, and Python shell commands as per the requirement.
- Developed Scala scripts, UDFs using both Data frames/SQL/Data sets and RDD/Map Reduce in Spark 1.6 for Data Aggregation, queries and writing data back into OLTP system through Sqoop.

- Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
- Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
- Experienced in querying HBase using **Impala**
- Developed Pig Scripts, Pig UDFs and Hive Scripts, Hive UDFs to analyze HDFS data.
- Extracted files from MongoDB through Sqoop and placed in HDFS and processed.
- Maintained the cluster securely using Kerberos and making the cluster up and running all times.
- Implemented optimization and performance testing and tuning of **Hive and Pig**.
- Experience in migrating HiveQL into Impala to minimize query response time.
- Developed a data pipeline using **Kafka** to store data into HDFS.
- Worked on reading multiple data formats on HDFS using **Scala**
- Worked on Spark SQL and Data frames for faster execution of Hive queries using PySpark **SqlContext**.
- Performed analysis on implementing PySpark using **Scala**.
- Implemented spark sample programs in python using **pyspark**.
- Designed and developed internal business systems/applications and took primary role in smaller, low risk projects
- Promoted and ensured compliance to the enterprise risk controls of the organization and external regulations
- Ensured Big Data practices integrate into overall data architectures and data management principles (e.g. data governance, data security, metadata, data quality)
- Assist in the development of comprehensive and strategic business cases used at management and executive levels for funding and scoping decisions on Big Data solutions
- Performance tuning of a Hadoop processes and applications

Environment: Snowflake**,** Openshift, PySpark, Impala, Scala, Kinesis, Shell Scripting, Eclipse, AWS, MySQL, Talend, hbase, Snowflake

Finning International **Aug 2021 – Jul 2022**

Big Data Application Developer

Project Summary : This project involves setting up a data repository and use it for search processing for analytical and research purposes. This application provides the capability for large batch processing using Hadoop map reduce jobs using Java runtime environment as well as real time search capabilities using Solr cloud environment using Pyspark

Responsibilities:

- Developed **Pyspark** programs to parse the raw data, and create intermediate data which would be further used to be loaded into Hive portioned data.
- Involved in creating **Hive** ORC tables, loading the data into it and writing Hive queries to analyze the data.
- Involved in data ingestion into HDFS using **Sqoop** for full load and **Flume** for incremental load on variety of sources like web server, RDBMS and Data API’s.
- Performed multiple MapReduce jobs in **PIG** and Hive for data cleaning and pre-processing
- Used different file formats like Text files, Sequence Files, Avro, Optimized Row Columnar (ORC)
- Ingest real-time and near-real time (NRT) streaming data into **HDFS** using Flume
- Expertise in creating TWS Jobs and Jobstreams and automate them as per schedule
- Worked on Golden Gate replication tool to get data from various data sources into HDFS

- Worked on **HBase** for support enterprise production and loading data into HBASE using **SQOOP**.
- Collecting and aggregating large amounts of log data using Apache Flume and staging data in HDFS for further analysis.
- Exported the data from Avro files and indexed the documents in ORC file format.
- Responsible for created Technical Specification documents for the generated extracts
- Involved in performance tuning using Partitioning, bucketing of **Hive** tables
- Created **UDFs** to calculate the pending payment for the given customer data based on last day of every month and used in **Hive** Scripts.
- Involved in writing shell scripts to run the jobs in parallel and increase the performance
- Involved in running TWS jobs for processing millions of records using ITG.

  Environment: Hadoop, HDFS, MapReduce, Yarn, Hive, PIG, Oozie, Sqoop, HBase, Flume, Linux, Shell scripting, Java, Eclipse, SQL
- Migrated an existing on-premises application to **AWS**. Used AWS services like EC2 and S3 for small data sets processing and storage, Experienced in Maintaining the Hadoop cluster on AWS **EMR**.
- Experience in **AWS EC2**, configuring the servers for Auto scaling and Elastic load balancing.
- Configuring AWS EC2 instances in VPC network & managing security through IAM and Monitoring servers' health through Cloud Watch.
- Responsible for creating, modifying topics (Kafka Queues) as and when required with varying configurations involving replication factors and partitions.
- Written shell scripts and **Python scripts** for automation of job.
- Providing problem analysis, viable solutions meeting the business needs in line with technology roadmap, set intermediate target
- dates, determine resources, track project completion & determine methods to accomplish goals
- Effectively Communicated with various Risk Areas and provided support for the Ingestion Tool in Environments (DEV,QA,PROD)
- Developed test approach & scripts for the project, coordinated system test activities.
- Used WebHdfs Java Api to move Data and Meta files from local file system to Hadoop file system

Enbridge **Jan 2017 – Jul 2021**

Big Data/Scala Developer

Project Summary: Designed and developed big data solutions involving Terabytes of data. The big data solution consists of collecting large amounts of log data from distributed sources, transformations and standardizations analysis, statistics, aggregations and reporting etc.

Responsibilities:

- Developed and optimized distributed applications using **Scala** and **Apache Spark** for big data processing.
- Designed and implemented high-performance data pipelines utilizing **Scala** and **Akka** for concurrent processing in real-time systems.
- Wrote clean, reusable, and efficient **Scala** code for data transformation and processing using **Spark SQL**, **Spark Streaming**, and **DataFrames** API.
- Built and maintained scalable data processing solutions using **Apache Spark**, **Hadoop**, and **Hive** for processing petabytes of data.
- Implemented real-time data streaming applications using **Apache Kafka** and **Spark Streaming** to process and analyze real-time data streams.

Bell Canada, Network Big Data COE **Apr 2016 – Dec 2016**

Big Data Developer

Project Description: Application is Customer Data Refinery. AWS hands-on experience with PySpark, Jenkins, Python, Docker. Serverless environment – all deployments developed in Docker and deployed in EC2 instances. Familiar with AWS EC2 instances, debugging logs.

Responsibilities:

- Hands-on technical knowledge of Bell Canada AWS environment, processes and procedures.
- Hands-on experience in Apache Spark, Jenkins, Python and Docker.
- Use a Cloud data warehouse such as Snowflake to query the data.
- Developed pipelines in **PySpark**, Python, Docker.
- Worked on enhancing existing pipeline built, Data Quality Checks in Snowflake Data Warehouse.
- Worked on External data such as Lexis-Nexis using the Rest API / Graph and used Data bricks Spark SQL for analysis.
- Worked with AWS CloudWatch Logs retention to be set as per the Log Ingestion Procedure.
- Worked with the AWS Quicksight report for making sure the S3 buckets are compliant in multi region.
- Used the Postman API to map the CNAME, Perform the Validations in CloudWaze to map the application to appropriate latency record in QA, PROD.
- Work on the Cloud Doctor to ensure the Auto Fail over feature is enabled for the data pipelines built in the mutli region.

RBC Capital Markets, RDARR **Jan 2016 – Mar 2016**

Big Data Application Developer

Project Description: Worked on building a generic utility to upload data with Hadoop. RDARR is a data project mainly built on hadoop using Java API's

Responsibilities:

- Hands-on technical knowledge of RBC RDARR Project requirements, processes and procedures.
- Hands-on experience in Apache Spark, Jenkins, Java and Docker.
- familiarity with data transformation and persistence in java to relational database SQLServer, MySQL
- Developed pipelines in Spark, Python, Docker.
- Familiarity with JSON processing using a 3 rd party library such as Jackson.
- Having experience on Hadoop eco system components HDFS**, MapReduce, Hive, Pig, Sqoop and Pyspark, HBase**.
- Worked with Logs retention to be set as per the Log Ingestion Procedure.
- Moved the data to HDFS and helped other teams access the data
- Deploy to Dev, QA, and PROD using CICD pipeline.
- Worked with cloud platforms like **AWS**, **Azure**, and **Google Cloud** to deploy scalable big data solutions leveraging **EMR**, **S3**, and **Databricks**.
- Set up and managed distributed data processing clusters using **YARN** and **Mesos** for large-scale data processing.
- Integrated cloud-based data lakes and warehouses (e.g., **Redshift**, **BigQuery**) with **Scala-based** data pipelines for fast querying and analytics.

Wolseley Canada **Jul 2013 – Dec 2015**

Data Analyst

Project Summary: As a Data Analyst, I have led and contributed to several high-impact projects that enabled data-driven decision-making and improved business processes. One of my major projects involved the **development of interactive dashboards** using **Tableau** and **Power BI**, which streamlined reporting and provided real-time insights for marketing, sales, and operations teams. These dashboards helped stakeholders track key performance indicators (KPIs) and business trends, ultimately improving decision-making speed and accuracy.

Responsibilities:

- Automated daily sales and marketing reporting using **SQL** and **Power BI**, saving the team **10+ hours** per week and improving reporting accuracy.
- Conducted a customer behavior analysis that led to a **20% increase** in targeted email campaign conversions.
- Streamlined inventory tracking, reducing stockouts by **15%** through data-driven optimization models.
- Created an interactive sales performance dashboard, enabling senior leadership to track KPIs in real time and improving decision-making speed.
- Interacting with the system analysts, business users for design & requirement clarifications.
- Designed front end pages using Jsp, HTML, Angular JS, JQuery, JavaScript, CSS and Ajax calls to get the required data from backend.
- Reduced data processing time by **30%** by implementing an automated data pipeline with **Python** and **dbt**.
- Created a dashboard in **Tableau** that streamlined reporting, saving the team **20 hours per week** in manual data compilation and analysis.
- Improved customer retention by **15%** through data-driven insights into subscription trends and user behavior.

Education:

GRADUATE CERTIFICATE in Information Technology and Web Design **May 2011-Apr 2013**
Lambton College of Applied Arts and Technology **Canada**

BACHELOR’S DEGREE in Computer Science and Engineering **Sep 2006-Apr 2010**
JNT University **India**