**Tumkesh Yadav**   

<tumkeshy037@gmail.com>

**Professional Summary:**

* Over 10+ years of diversified IT experience in E2E data analytics platforms (ETL - BI-Java) as Bigdata, Hadoop, Java/J2EE Development, Informatica, Data Modeling and System Analysis.
* Hands on experience Hadoop framework and its ecosystem like Distributed file system (HDFS), MapReduce, Pig, Hive, Sqoop, Flume, Spark.
* Experience in layers of Hadoop Framework - Storage (HDFS), Analysis (Pig and Hive), Engineering (Jobs and Workflows), extending the functionality by writing custom UDFs.
* Extensive experience in developing Data warehouse applications using Hadoop, Informatica, Oracle, Teradata, MS SQL server on UNIX and Windows platforms and experience in creating complex mappings using various transformations and developing strategies for Extraction, Transformation and Loading (ETL) mechanism by using Informatica.
* Experience with building data pipelines in Python/Pyspark/HiveSQL/Presto/BigQuery and building python DAG in Apache Airflow.
* Proficient in Hive Query language and experienced in hive performance optimization using Static-Partitioning, Dynamic-Partitioning, Bucketing and Parallel Execution concepts.
* Experience in analyzing data using Hive QL, Pig Latin, and custom MapReduce programs in Java, custom UDF s.
* Good Understanding of Hadoop Architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node and MapReduce concepts.
* Knowledge on Cloud computing infrastructure AWS (amazon web services).
* Expertise in creating modules for spark streaming in data into Data Lake using Spark.
* Experience in Dimensional Data Modeling Star Schema, Snow-Flake Schema, Fact and Dimensional Tables, concepts like Lambda Architecture, and Batch processing, Oozie.
* Good experience in Informatica client tools Source Analyzer, Warehouse designer, Mapping designer, Mapplet Designer, ETL Transformations, Informatica Repository Manager and Informatica Server Manager, Workflow Manager & Workflow Monitor.
* Experience with Snowflake cloud data warehouse and AWS S3 bucket for int egrating data from multiple source system which include loading nested JSON formatted data into snowflake table.
* Experience in writing SQL, PL/SQL queries, Stored Procedures for accessing and managing databases such as Oracle, AnsiSQL, PostgreSQL, MySQL, TSQL.
* Exposure to AI and Deep learning platforms such as TensorFlow, Keras, AWS ML, Azure ML studio
* Experience on Migrating SQL database to Azure data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse and Controlling and granting database access and Migrating On premise databases to Azure Data lake store using Azure Data factory
* Expertise in using core Java, J2EE, Multithreading, JDBC, Shell Scripting and proficient in using Java API's Collections, Servlets, JSP for application development.
* Experience in reviewing pre- and post-processed data to ensure data accuracy and integrity along with Dev and QA teams.
* Experience in Java, J2ee, JDBC, Collections, Servlets, JSP, Struts, Spring, Hibernate, JSON, XML, REST, SOAP Web services, Groovy, MVC, Eclipse, Weblogic, Websphere, and Apache Tomcat severs.
* Strong experience with Functional programming languages like Scala, and Java.
* Extensive knowledge of Data Modeling, Data Conversions, Data integration and Data Migration with specialization in Informatica Power Center.
* Expertise in extraction, transformation and loading data from heterogeneous systems like flat files, excel, Oracle, Teradata, MSSQL Server.
* Good experience in UNIX/Linux commands, scripting and deploying the applications on the servers.
* Strong skills in algorithms, data structures, Object oriented design, Design patterns, documentation and QA/testing.
* Experience in working as part of fast paced Agile Teams, exposure to testing in scrum teams,Test-Driven development.

**Skills:**

|  |  |
| --- | --- |
| **BigData Technologies** | AWS EMR, S3, EC2-Fleet, Spark-2.2, 2.0 and 1.6, Hortonworks HDP, Hadoop, Mapreduce, Apache Spark, SparkSQL, Informatica Power Center 9.6.1/8.x, Kafka, NoSQL, Elastic Mapreduce(EMR), Sqoop, Solr. |
| **Databases** | Cloudera Hadoop CDH 15.x, Hortonworks HDP, Oracle 10g/11g, Teradata, DB2,Microsoft SQL Server, MySQL, noSQL,SQL databases. |
| **Platforms (O/S)** | Red-Hat LINUX, Ubuntu, Windows NT/2000/XP. |
| **Programming languages** | Java, Scala, SQL, UNIX shell script, JDBC, Python, Perl. |
| **Security Management** | Hortonworks Ambari, Cloudera Manager, Apache Knox, XA Secure, Kerberos . |
| **Web-technologies** | HTML, XML, XSL (XSLT, XPATH), XSD, CSS, JavaScript, SOAP, RESTful, Agile. |
| **Data warehousing** | Informatica Powercenter, Pentaho, ETL Development, Amazon Redshift, IDQ. |
| **Database Tools** | JDBC, HADOOP, Hive, No-SQL, SQL Navigator, SQL Developer, TOAD, SQL Plus, SAP Business Objects, ANSI-SQL |
| **Data Modeling** | Rational Rose, Erwin 7.3/7.1/4.1/4.0 |
| **Code Editors** | Eclipse, Intellij |

**Professional Experience:**

**Lead Data Engineer**

**Deutsche bank(Remote)**

**April 2022 – Current**

**Client Description-**

Deutsche Bank AG (DB) is a provider of investment banking, corporate and retail banking, and asset and wealth management solutions. Its portfolio of offerings comprises deposit services, card products, loans, trust, and securities solutions, wealth, and investment planning, financial advisory, and financing.

**Project Description-**

The project focuses on developing a scalable data engineering and analytics platform that leverages Big Data technologies to process and analyze large and diverse datasets. As the Lead Data Engineer, I had played a crucial role in designing, developing, and maintaining the data pipelines and infrastructure, collaborating with cross-functional teams, and ensuring the quality and integrity of the data used in ML applications

**Responsibilities:**

* Leveraged Spark RDDs to process and analyze data by extracting it from HDFS, enabling predictive analytics capabilities.
* Utilized Hive Context to access data, leveraging the advanced functionality offered by SQLContext. Employed HiveQL queries for data retrieval from Hive tables, including fact and syndicate tables.
* Employed efficient data modeling techniques in Hive, extensively utilizing partitioning for enhanced data separation and faster processing. Followed industry best practices to optimize Hive performance.
* Developed custom Scala RDDs to perform complex data transformations and execute actions on RDDs within Spark scripts. Implemented caching of RDDs to improve performance and facilitate actions on each RDD.
* Designed and constructed Hive Fact tables to process raw data from multiple retailers. These tables were partitioned based on the Time dimension key, Retailer name, and Data supplier name. These tables served as inputs to analytics service engines.
* Develop stored procedures/views in Snowflake and use in Talend for loading Dimensions and Facts.
* Developed and maintained data pipelines using Snowpark and SnowPipe in Snowflake.
* Leveraged AWS infrastructure, including AWS EMR Distribution for Hadoop, AWS S3 for raw file storage, and AWS EC2 for Kafka.
* Utilized AWS Lambda for performing data validation, filtering, sorting, and other transformations triggered by data
* Developed highly maintainable and user-friendly Python and Scala code, meeting application requirements for data processing and analytics. Utilized built-in libraries to facilitate efficient development.
* Designing, developing, and maintaining data pipelines using Snowpark and SnowPipe in Snowflake.
* Monitoring Big query, Dataproc and cloud Data flow jobs via Stack driver for all the different environments.
* Performed thorough data analysis for the purpose of overhauling the database using ANSI-SQL.
* Developed UDF’s in Java as and when necessary to use in Hive queries.
* Build the Logical and Physical data model for snowflake as per the changes required
* Designed and developed Microservices business components using Spring Boot.
* Build a program with Python and apache beam and execute it in Cloud Dataflow to run Data validation between raw source file and Bigquery tables.
* Evaluated Snowflake Design considerations for any change in the application
* Coded process using PLSql, used dynamic views using parameterized user environment variables.
* Designing and implementing scalable and efficient data pipelines, managing and maintaining large datasets, collaborating with data scientists and ML engineers to develop and deploy ML models, and ensuring the quality and integrity of data used in ML applications.
* Extracted data from various APIs, data cleansing and processing by using Java and Scala
* Writing PLSQL Packages to Extract, Transform and Load data into Target tables
* Implemented scalable microservices applications using Spring boot for building Rest end points.
* Led the development of scalable distributed data solutions within Amazon EMR cluster environments. Leveraged EMR cluster capabilities to build robust data processing systems.
* Utilized Kafka REST API to collect and load data onto the Hadoop file system. Incorporated Sqoop for seamless data loading from relational databases.
* Extracted real-time data feeds using Kafka and Spark Streaming. Converted the data into RDD format, processed it using DataFrames, and stored the results in Parquet format within HDFS.
* Successfully loaded files from Oracle and SQL Server into Hive and HDFS using Sqoop, ensuring smooth data integration across systems.
* Prepared complex ANSI-SQL queries, views and stored procedures to load data into staging area.
* Implemented data quality checks and data validation processes to ensure the accuracy and reliability of the processed data.
* Collaborated with cross-functional teams to understand data requirements and provide data engineering solutions aligned with business needs.
* Optimized data pipelines by implementing data partitioning, compression techniques, and efficient data serialization formats like Avro and Parquet.
* Loading the data from multiple Data sources like (SQL, DB2, and Oracle) into HDFS using Sqoop and load into Hive tables.
* Integrated Spark with other data processing frameworks such as Apache Kafka, Apache HBase, and Apache Cassandra to enable seamless data ingestion and real-time analytics capabilities.
* Migrate data into RV Data Pipeline using DataBricks, Spark SQL and Scala.
* Conducted performance tuning and optimization of Spark jobs, analyzing resource utilization, and fine-tuning configuration parameters for enhanced job execution efficiency.
* Implemented continuous integration and deployment (CI/CD) pipelines using tools like Jenkins and Docker to automate build, testing, and deployment processes.
* Actively contributed to code reviews, conducted unit testing, and implemented monitoring and logging solutions to ensure the stability and reliability of data pipelines.
* Written spark Application to capture the change feed from the DocumentDB using java API and write updates to the new DocumentDB.
* Used Databricks for encrypting data using server-side encryption.
* Worked on scaling microservices applications using Kubernetes and docker.

**Environment:** Cloudera Hadoop , Snowflake, AWS EMR,AWS S3, AWS Glue, ANSI-SQL, PLSQL, TSQL Big Query, SQOOP, Data Factory, Data Bricks, Spark-Java, Spark-Scala, Hive, Impala, HDFS, Oozie, TFS, Agile, MS-SQL, PySpark, Eclipse, Kafka, Sqoop, Spark Streaming, Parquet, Avro, Apache HBase, Apache Cassandra, Redshift, Jenkins.

**Sr. Data Engineer**

**Nike, New York**

**Dec 2019-March 2022**

**Client Description-**

Nike, Inc. is an American multinational association that is involved in the design, development, manufacturing and worldwide marketing and sales of apparel, footwear, accessories, equipment and services. The company's world headquarters are situated near Beaverton, Oregon, in the Portland metropolitan area (USA).

**Project Description-**

As a Data Engineer, I was responsible for designing, implementing, and coding various components of a data system. I also played a role in developing a framework to generate daily adhoc reports and extracts from enterprise data using BigQuery. This involved developing complex SQL queries, including stored procedures, common table expressions (CTEs), and temporary tables, following ANSI SQL standards. The reports were utilized for analysis and visualization using Power BI and SSRS.

**Responsibilities:**

* Involved in design, implementation and coding in XML, Java Servlets, JSP, and JavaScript.
* Built scalable distributed data solutions in an EMR cluster environment with Amazon EMR 5.6.1.
* Utilized Kafka REST API and Sqoop to collect and load data into the Hadoop file system from both streaming and relational databases.
* Extracted real-time data feeds using Kafka and Spark Streaming, processed the data as RDDs and DataFrames, and stored it in Parquet format in HDFS.
* Utilized Spark Streaming APIs to perform necessary data transformations and actions on the data received from Kafka, persisting the results in HDFS.
* Coordinated with team and Developed framework to generate Daily adhoc reports and Extracts from enterprise data from BigQuery.
* Developed complex SQL queries using stored procedures, common table expressions (CTEs), temporary table to supportPower BIand SSRS reports uning ANSI SQL standards.
* Developed MapReduce jobs in java for data cleaning and preprocessing
* Implemented multiple modules in microservices to expose data through Restful Api’s.
* Developed custom Spark scripts in Scala, leveraging RDDs for efficient data transformations and actions.
* Leveraged Hive for providing an abstraction layer on top of structured data in HDFS, implementing various techniques such as Partitions, Dynamic Partitions, and Buckets on HIVE tables.
* Designed and implemented a fully operational production grade large scale data solution on Snowflake Data Warehouse.
* Developed GUI HTML, XHTML, AJAX, CSS 5 and JavaScript (jQuery) and exclusively used CSS for modifying Layout and design of the web pages.
* Developed different modules in microservices to collect stats of application for visualization.
* Utilized Spark API with Hadoop YARN as the execution engine for data analytics, leveraging the power of Hive.
* Performed advanced text analytics and processing using Spark's in-memory computing capabilities, utilizing Scala.
* Migrated MapReduce programs into Spark transformations using Scala.
* Designed and developed data integration programs in a Hadoop environment, incorporating NoSQL data store Cassandra for data access and analysis.
* Utilized Apache Oozie as the job management scheduler to execute workflows.
* Monitored node health and job status in Hadoop clusters using Ambari.
* Designed and implemented data warehouses and data marts following Kimball Methodology components such as Data Warehouse Bus, Conformed Facts & Dimensions, Slowly Changing Dimensions, Surrogate Keys, Star Schema, Snowflake Schema, etc.
* Performed information purging and applied changes utilizing Databricks and Spark information analysis.
* Utilized Databricks notebooks for interactive analysis utilizing Spark APIs.
* Utilized Tableau to build customized interactive reports, worksheets, and dashboards.
* Implemented LDAP and Active Directory for Hadoop clusters.
* Utilized Apache Solr for indexing and load-balanced querying in larger datasets.
* Performed performance tuning of Spark jobs, utilizing caching and maximizing the advantages of the cluster environment.
* Created Spring-Boot services for Oozie orchestration and deployed entity services for the Audit Framework of the loaded data.
* Worked related to downloading BigQuery data into pandas or Spark data frames for advanced ETL capabilities.
* Created Databricks notebooks using ANSI SQL, Python and automated notebooks using jobs.
* Worked on plsql performance, I had improved teh coding standards by using my past experience.
* Worked with SQL\* Loader, PLSQL Packages for Data Extraction from Legacy to Oracle Database.
* Used AWS S3 to store large amounts of data in a reliable repository

**Environment:** EMR, Amazon EMR, Snowflake, AWS S3, AWS Glue ,ANSI-SQL, PLSQL,TSQL, ANSI-SQL, Kafka, Data Factory, Data Bricks, Sqoop, Spark Streaming, Scala, Spring Boot, Oozie, Avro, Parquet, ORC, Hive, HDFS, YARN, Cassandra, Ambari, Tableau, Kerberos, LDAP, Apache Solr.

**Data Engineer**

**QBurst,** **San Francisco, CA**

**Nov2018 - Nov 2019**

**Client Description-**

QBurst is a global product development and consulting company offering cognitive solutions and custom software development services for micro to large enterprises

**Project Description-**

The project involved designing and implementing an efficient Extract, Transform, and Load (ETL) process for handling diverse data feeds in a Data Lake environment

**Responsibilities:**

* Involved in testing Snowflake to understand best possible way to use the cloud resources.
* Prepared comprehensive ETL design documents encompassing database structure, change data capture, error handling, restart and refresh strategies.
* Implemented Azure Data Factory (ADF) extensively for ingesting data from different source systems like relational and unstructured data to meet business functional requirements
* Worked with diverse data feeds including JSON, CSV, XML, and DAT, implementing the Data Lake concept.
* changes in a DynamoDB table. Transformed data was loaded into another data store.
* Programmed ETL functions for data integration between Oracle and Amazon Redshift.
* Design Setup maintain Administrator the Azure SQL Database, Azure Analysis Service, Azure SQL Data warehouse, Azure Data Factory, Azure SQL Data warehouse.
* Built interactive and dynamic action trigger platform using Python, Azure ML to notify trial site for potential abnormal behavior
* Built a common sftp download or upload framework using Azure Data Factory and Databricks.
* Developed custom alerts using Azure Data Factory, SQLDB and Logic App.
* Gained exposure to IRI's end-to-end analytics service engine and new big data platforms such as the Hadoop loader framework and Big Data Spark framework.
* Utilized Kafka producer to ingest raw data into Kafka topics and ran Spark Streaming applications to process clickstream events.
* Conducted data analysis and implemented predictive data modeling techniques.
* Utilized Spark SQL within the Apache Spark big data framework to explore clickstream event data and perform structured data processing for shipment, POS, consumer, household, and individual digital impressions.
* Loaded terabytes of raw data into Spark RDDs for data computation and generating output responses.
* Led a significant initiative focused on media analytics and forecasting, aiming to deliver sales lift associated with customer marketing campaign initiatives.
* Leveraged Python for automation of Hive operations and reading configuration files.
* Utilized Spark for high-speed data processing, leveraging both Spark Shell and Spark Standalone cluster environments.
* Leveraged Hive for analyzing partitioned data and computing various metrics for reporting purposes.
* Creation, configuration and monitoring Shards sets. Analysis of the data to be shared, choosing a shard Key to distribute data evenly. Architecture and Capacity planning for MongoDB clusters. Implemented scripts for mongo DB import, export, dump and restore.
* Write ANSI SQL to develop complex transformations and derivations
* Used MongoDB internal tools like Mongo Compass, Mongo Atlas Manager & Ops Manager, Cloud Manager etc.
* Worked on MongoDB database concepts such as locking, transactions, indexes, sharing, replication and schema design.
* Worked on designing and developing Multi-tier enterprise level web applications using various J2EE technologies including Servlets2.x, JDBC, Apache Ant1.5, HTML, XHTML, DHTML, CSS, JavaScript3.x, JSP and XML technologies.
* Created and deployed ETL mappings and develop PL SQL procedures to execute complex data transformations
* Used Azure Key vault as central repository for maintaining secrets and referenced the secrets in Azure Data Factory and also in Databricks notebooks.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Created DataFrames from various data sources, including existing RDDs, structured data files, JSON datasets, Hive tables, and external databases.

**Environment:** Map Reduce, ADF, Azure, Snowflake, , ANSI-SQL, PLSQL, TSQL, HDFS, Hive, Python, Scala, Kafka, Spark, Spark Sql, Oracle, Informatica 9.6, SQL, MapR, Sqoop, Zookeeper, ,Data Pipeline, Jenkins, GIT, JIRA, Unix/Linux, Agile Methodology, Scrum.

**Data Engineer**

**Next Gen Healthcare, NY**

**September 2016-October 2018**

**Client Description-**

ACL Digital is a leading digital marketing agency in India, offering a range of services including SEO, PPC, social media marketing, and content marketing

**Project Description-**

This project focused on designing and implementing a Big Data infrastructure for processing and analyzing large datasets. I had Provided support for running MapReduce Java programs on the cluster. This involved troubleshooting and optimizing the performance of MapReduce jobs.

**Responsibilities:**

* Understood project requirements and prepared architecture documentation for the Big Data project.
* Supported MapReduce Java programs running on the cluster.
* Optimized Amazon Redshift clusters, Apache Hadoop clusters, data distribution, and data processing.
* Developed MapReduce programs to process Avro files, performing calculations on the data and implementing map-side joins.
* Responsible for Migration of key systems from on-premises hosting to Azure Cloud Services Snow SQL Writing: SQL queries against Snowflake.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the Sql Activity.
* Imported bulk data into HBase using MapReduce programs.
* Utilized REST API to access HBase data for analytics purposes.
* Designed and implemented incremental imports into Hive tables.
* Created Hive tables, loaded them with data, and wrote Hive queries running internally in a MapReduce manner.
* Collected, aggregated, and moved data from servers to HDFS using Flume.
* Imported and exported data from various relational data sources (DB2, SQL Server, Teradata) to HDFS using Sqoop.
* Built applications using Maven and integrated with continuous integration servers like Jenkins.
* Leveraged an Enterprise Data Warehouse database to store information and make it accessible across the organization.
* Responsible for preparing technical specifications, analyzing functional specifications, and developing and maintaining code.
* Developed Snowflake views to load and unload data from and to an AWS S3 bucket, as well as transferring the code to production.
* Analyze Data and Performed Data Preparation by applying historical model on the data set in AZURE ML.
* Collaborated with the Data Science team to gather requirements for various data mining projects.
* 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.
* Developed shell scripts for automating day-to-day processes.
* Migrated complex MapReduce programs to in-memory Spark processing using transformations and actions.
* Conducted a proof of concept (POC) for IoT device data using Spark.
* Utilized Scala to store streaming data in HDFS and implement Spark for faster data processing.
* Created RDDs and DataFrames for input data and performed data transformations using Spark Python.
* Developed Spark SQL queries, DataFrames, imported data from various sources, performed transformations, read/write operations, and saved results to the output directory in HDFS.
* Implemented Hive jobs to parse and structure log data for effective querying.
* Developed PIG scripts for the analysis of semi-structured data.
* Created PIG UDFs for manipulating data according to business requirements and developed custom PIG loaders.
* Utilized Oozie workflow engine for job scheduling and developed Oozie workflows for scheduling and orchestrating ETL processes.
* Managed and reviewed Hadoop log files using shell scripts.
* Migrated ETL jobs to Pig scripts for transformations, joins, and pre-aggregations before storing data in HDFS.
* Worked with different file formats such as Sequence files, XML files, and Map files using MapReduce programs.
* Utilized Avro Data Serialization system to work with JSON data formats.
* Worked with the HortonWorks distribution of Hadoop.

**Environment:** HortonWorks, ADF, Snowflake, Azure, MapReduce, Amazon Redshift, Apache Hadoop, Avro, HBase, REST API, Hive, Flume, Sqoop, Spark, Scala, Python, Spark SQL, PIG, Oozie, Shell scripting, XML, Maven, Jenkins.

**Software Engineer**

**Verizon,NY**

**June 2013 – August 2016**

**Client Description-**

Verizon is a multinational telecommunications company headquartered in the United States. It provides wireless, broadband, and telephony services to customers in the United States, Canada, and internationally  
**Project Description-**

This project revolved around developing ETL processes to extract data from ERP systems, Oracle databases, and other sources. The extracted data was transformed using Informatica's SQ, Router, Expression, Aggregator, Lookup, and other transformations. The transformed data was then loaded into both the target Hadoop Distributed File System (HDFS) environment and a Teradata database.

**Responsibilities:**

* Involved in developing ETL processes to extract data from ERPs and Oracle databases, transform it using Informatica's SQ, Router, Expression, Aggregator, Lookup, and other transformations, and load it into the target HDFS environment and Teradata database. Also, handled metadata in MySQL for data loading into HDFS.
* Do Participated in architecture council for database architecture recommendation
* Utilized Unix Shell Scripts for adding the header to the flat file targets.
* Used Teradata utilities fastload, multiload, tpump to load data.
* Preparation of the Test Cases and involvement in Unit Testing and System Integration Testing.
* Deep analysis on SQL execution plan and recommend hints or restructure or introduce index or materialized view for better performance
* Utilized Power Query in Power BI to Pivot and Un-pivot the data model for data cleansing.
* Used ASP.NET Caching to improve the performance of the Application.

**Environment:** Informatica, Oracle, Teradata, MySQL, Unix Shell scripting, PL/SQL, MapReduce, HDFS, Pig, Hive.  
  
**Bachelors Details-**

**Bachelors in Computer Science, Rajeev Gandhi Prodhyogik Vishwavidhalaya, 2013**

**Certifications-**

* **Certification of ALEKS- MSBIA Statistics proficiency.**
* **Certification of completion HTML, PHP, CSS, WordPress, Java Script.**
* **Certification of completion C and C++.**