**Lava Kumar**

**reddylava840@gmail.com | +1 (732) 554-6395**

# Professional Summary

* **Having around 10 years** of professional experience in IT, working with various Legacy Database systems, which include work experience in Big Data technologies as well.
* Good experience on understanding of architecting, designing and operation of large scale data and analytics solutions on **Snowﬂake Cloud Data Warehouse**.
* Experience in **Requirement gathering, System analysis, handling business** and technical issues & communicating with both business and technical users.
* Hands on experience on complete **Software Development Life Cycle SDLC** for the projects using methodologies like

**Agile** and hybrid methods.Strong experience in migrating other databases to **Snowflake**.

* Experience in analyzing data using **Big Data Ecosystem** including **HDFS, Hive, HBase, Zookeeper, PIG, Sqoop, and Flume.**
* Designed and implemented complex ETL processes to transform and load data from various sources into Snowflake, ensuring data consistency and accuracy.
* Knowledge and working experience on big data tools like **Hadoop**, **Azure Data Lake**, **AWS Redshift**.Good understanding of **Apache Airﬂow**.
* Hands-on experience in using NOSQL libraries like MongoDB, Cassandra, and Redis and relational databases like Oracle, SQLite, PostgreSQL, and MYSQL databases.
* Managing Database, **Azure Data Platform** services (**Azure Data Lake (ADLS**), **Data Factory (ADF), Data Lake Analytics**, **Stream Analytics, Azure SQL DW, HDInsight/Data bricks, NoSQL DB), SQL Server, Oracle, Data Warehouse etc. Build multiple Data Lakes.**
* Used Couchbase Lite, a client-side **‘NoSQL’** database hosted on the device, exposed via **REST API** through the domain which automatically replicates to a master server-side database.
* Worked on **Rest Web services** for backend services, used **Couchbase DB (NoSQL)** for database services.
* services.
* Hands-on experience in **using NOSQL libraries like MongoDB, Cassandra, and Redis and relational databases like Oracle, SQLite, PostgreSQL, and MYSQL databases.**
* Experience in workﬂow scheduling with **Airﬂow, AWS Data Pipelines, Azure, SSIS, etc.**
* 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**.
* Good understanding of **Big Data Hadoop** and **Yarn architecture** along with various **Hadoop Demons** such as **Job Tracker, Task Tracker, Name Node, Data Node, Resource/Cluster Manager, and Ka a (distributed stream- processing).**
* Conducted training sessions for the team to enhance Snowflake proficiency.
* Experience in Text Analytics, **Data Mining** solutions to various business problems and generating **data visualizations**

using **SAS** and **Python**.Consulting on **Snowflake Data Platform** Solution Architecture, Design, Development and deployment focused to bring the data driven culture across the enterprises

* Experience in Developing **Spark applications** using **Spark - SQL in Databricks** for **data extraction**, transformation and aggregation from multiple ﬁle formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Good understanding of **Spark** Architecture including **Spark Core, Spark SQL, Data Frames, Spark Streaming, Driver Node, Worker Node, Stages, Executors** and **Tasks**.
* Implemented security best practices in Azure Databricks, including authentication, authorization, and data encryption, ensuring data security and compliance with Azure data governance standards.
* Collaborated with cross-functional teams to design and implement data solutions using Azure Databricks, showcasing teamwork and communication skills in a collaborative Azure environment.
* Assisted in building real-time data processing solutions using Apache Flink and Apache Kafka, gaining hands-on experience in stream processing and event-driven architecture.
* Stayed up-to-date with the latest Azure Databricks features, updates, and best practices, continuously improving skills and knowledge in Azure Databricks and data processing technologies in the Azure ecosystem.
* Experienced in processing real-time data using Kafka 0.10.1 producers and stream processors and implemented stream process using Kinesis and data landed into Datalake S3.
* Involved in ETL Data validations of packages, data Analytics and various business rules modifications based on CMS guidelines and verify the Business transactions flow across multiple systems, migrated 3 modules from SQL DB to Azure data lake, Azure SQL DB and Data Bricks using ADF.
* Strong experience and knowledge of **NoSQL databases** such as **MongoDB** and **Cassandra**.
* Experience in development and support knowledge on **Oracle, SQL, PL/SQL, T-SQL queries.**
* Experienced in **conﬁguring and administering** the **Hadoop Cluster** using major **Hadoop** Distributions like **Apache Hadoop and Cloudera**.
* Solid Excellent experience in creating cloud based solutions and architecture using **Amazon Web services (Amazon EC2, Amazon S3, Amazon RDS,EMR, Glue)** and **Microsoft Azure.**
* Experienced in Technical consulting and end-to-end delivery with **architecture**, **data modeling**, **data governance** and

**design** - development - implementation of solutions.

* Experience in **Big Data Hadoop Ecosystem** in ingestion, storage, querying, processing and analysis of **big data.**
* Extensive working experience in agile environment using a **CI/CD model.**
* Extensive experience working with structured data using **Spark SQL, Data frames, Hive QL,** optimizing queries, and incorporate complex **UDF's** in business logic.

# Technical Skills

* **Big Data & Hadoop Ecosystem** : Hadoop 3.3/3.0, Hive 2.3, Solr 7.2, Apache Flume 1.8, Sqoop 1.4, Ka a 1.0.1, Oozie 4.3, Hue, Cloudera Manager, Stream sets
* **Cloud Technologies**: AWS, Glue,EC2, EC3,EMR, Redshift & MS Azure, Snowﬂake
* **Data Modeling Tools**: Erwin R9.7, ER Studio v16,Snowflake
* **Packages**: Microsoft Oﬃce 2019, Microsoft Project, SAP and Microsoft Visio 2019, Share point Portal Serve

**Other Tools**: VSS, SVN, CVS. Docker, CI/CD, Kubernetes

* **RDBMS / NoSQL Databases:** Oracle 19c, Teradata R15, MS SQL Server 2019, Cosmos DB,Cassandra 3.11, HBase 1.2,
* **Testing and defect tracking Tools:** HP/Mercury, Quality Center, Win Runner, MS Visio 2016 & Visual Source Safe
* **Operating System:** Windows 10/8, Unix, Sun Solaris
* **ETL/Data warehouse Tools:** Informatica 9.6, SAP Business Objects XIR3.1/XIR2, Talend, Tableau
* **Methodologies:** RAD, JAD, RUP, UML, System Development Life Cycle (SDLC), Agile, Waterfall Model.

# BNSF- Fort Worth, TX Jan 2022 - Till Date

**Lead Data Engineer**

# Responsibilities:

* Involved in building database Model and Views utilizing **Python**, to build an interactive web based solution.
* Collaborated with other developers to handle complicated issues related with deployment of Django based applications.
* Handled development and management of front end user interfaces with help of **HTML**, **CSS**, **JQuery** and **JavaScript**.
* Modify the existing **Python/Django** modules to deliver certain format of data and adding new features.
* Consulting on **Snowflake Data Platform** Solution Architecture, Design, Development and deployment focused to bring the data driven culture across the enterprises
* Good work experience in RDBMS such as MySQL and SQL server and NoSQL Database like Mongo DB, Cassandra**.**
* Expert in writing SQL queries and using Hibernate framework with Spring ORM in interaction with the RDBMS and familiar with Non-Relational Database (NoSQL) like Mongo DB and Cassandra.
* Working with NoSQL databases like MongoDB to store and retrieve data.
* Strong experience in working with NoSQL databases like MongoDB.
* Automated a reporting process, using **Python**, **Luigi** (library for task workﬂow and dependencies), and other API’s.
* Written python scripts using **python** libraries such as **pandas**, numpy that does read/write operations on large CSV ﬁles, perform data aggregations and compare data by columns.
* Led the migration of on-premises data warehouses to Snowflake, resulting in a 30% reduction in operational costs and improved query performance.Created data models, schemas, and views in Snowflake to support business intelligence initiatives and decision-making.
* Experience in using **NoSQL** databases **Redis**.
* Dedicated and highly skilled Data Engineer with a proven track record in designing, developing, and maintaining data pipelines using Apache Flink.
* Successfully led the migration of a legacy data warehouse to Snowflake, including data modeling and ETL processes.
* Experience in integrating **Python REST API** Frame work using **Django**.
* Contributed to the design and development of real-time fraud detection system using Apache Flink, reducing fraudulent transactions by 25%.
* Working experience of **Data Warehouse ETL /design** and implementation of complex big data pipelines.
* Used **Python**, **PySpark**, shell script, **oracle** scheduler, **Luigi**, **Oracle Pl SQL** etc.
* Developed JSON scripts for deploying pipelines in Azure Data Factory (ADF) that process data using Azure Synapse SQL Activity, demonstrating experience with data pipeline development and integration with Azure services.
* Used Azure Databricks for processing large-scale data using distributed computing, including data ingestion, data transformation, and data analysis tasks, showcasing proﬁciency in utilizing Azure Databricks for big data processing
* Developed and implemented machine learning models using Azure Databricks, leveraging its built-in machine learning libraries and distributed computing capabilities, demonstrating expertise in machine learning on the Azure platform.
* Leveraged Azure Databricks notebooks for interactive data exploration, visualization, and analysis, showcasing proﬁciency in utilizing the collaborative notebook environment for data analysis and exploration.
* Integrated Azure Databricks with other Azure services, such as Azure Synapse Analytics, Azure Blob storage, and Azure SQL Database, for seamless data processing and analysis workﬂows, demonstrating proﬁciency in building end-to-end data solutions using Azure Databricks.
* Designed and implemented real-time data processing pipelines using Apache Flink, ensuring low-latency data ingestion and processing for critical business applications.
* Developed data connectors to integrate Apache Flink with various data sources and sinks, including databases, cloud storage, and third-party APIs.
* Collaborated with the DevOps team to automate deployment and monitoring of Snowflake environments.
* Managed and maintained data pipelines, integrating data from multiple sources into Snowflake.
* Utilized Azure Databricks for real-time data processing and stream analytics, showcasing expertise in processing and analyzing data in real-time using Azure Databricks streaming capabilities.
* Used Jenkins to deploy our code into diﬀerent environments and scheduling jobs.
* Used bug-tracking tools like Jira, conﬂuence and version controls Git, GitLab.

**Environment**: Python, Django, Luigi, windows, Linux, MySQL, SQL, Cassandra, AWS RDS, AWS S3, AWS EC2, Ka a, JSON, Restful API, MVC architecture, GitLab, Agile, Enterprise Scheduler, Bitvise SSH Client, Scrum, JIRA, GIT.

# ValueLabs-India Jul 2018- Feb 2021

**Data Engineer**

# Responsibilities:

* As a **Data Engineer** I am responsible for building scalable distributed data solutions using **Hadoop**.
* Involved in **Agile** Development process **(Scrum and Sprint planning).**
* Handled **Hadoop cluster** installations in **Windows** environment.
* Migrated on-premises environment in **GCP (Google Cloud Platform)**
* Migrated data warehouses to **Snowﬂake Data warehouse.**
* Implemented monitoring and alerting solutions for Apache Flink jobs, using tools like Prometheus and Grafana, to proactively identify and address issues.
* Frameworks for distributed data processing, including **Apache Spark** and **Apache Flink**, have been successfully installed on **GKE**, allowing for effective parallel processing of large datasets.
* Deﬁned **virtual warehouse** sizing for **Snowﬂake** for diﬀerent type of workloads.
* Integrated and automated data workloads to **Snowﬂake Warehouse.**
* Created tables in **snowﬂake DB,** loading and analyzing data using **Spark-Scala** scripts.
* Developed **ETL pipelines** in and out of data warehouse using combination of **Python** and **Snowﬂake’s Snow SQL.**
* Written **POCs** in **Python** to analyze the data quickly before applying **big data solutions** to process at a scale.
* Responsible for **data governance** rules and standards to maintain the consistency of the business element names in the

diﬀerent data layers.

* Build **Data Warehouse** in **Azure platform** using **Azure data bricks** and **data factory.**
* Developed **data pipeline** using **Sqoop** to ingest cargo data and customer histories into **HDFS** for **analysis**.
* Designed **ETL** using Internal/External tables and store in parquet format for eﬃciency.
* Involved in porting the existing on-premises **Hive code migration** to **GCP (Google Cloud Platform) BigQuery**.
* Involved in migration an **Oracle SQL ETL** to run on **Google cloud platform** using cloud **Dataproc** & **BigQuery**, cloud pub/sub for triggering the **Apache Airﬂow** jobs.
* Extracted data from **data lakes, EDW** to relational databases for analyzing and getting more meaningful insights using

# SQL Queries and PySpark.

* Used **AWS** services like **EC2** and **S3** for small data sets processing and storage, Experienced in Maintaining the **Hadoop**

cluster on **AWS EMR**.

* Designed, developed and did maintenance of **data integration** programs in a **Hadoop** and **RDBMS** environment with both traditional and non-traditional source systems.
* Developed **MapReduce** programs to parse the raw data, populate staging tables and store the reﬁned data in partitioned tables in the **EDW.**
* Wrote **Sqoop Scripts** for importing and exporting data from **RDBMS** to **HDFS**.
* Set up **Data Lake** in **Google cloud** using Google cloud storage, **BigQuery** and **Big Table.**
* Developed scripts in **BigQuery** and connecting it to reporting tools.
* Designed and implemented real-time data processing pipelines with Snowflake, enabling near-instant access to critical business data.
* Designed workﬂows using **Airﬂow** to automate the services developed for Change data capture.
* Carried out data transformation and cleansing using **SQL queries** and **PySpark**.
* Used **Ka a** and **Spark** streaming to ingest real time or near real time data in **HDFS**.
* Worked related to downloading **BigQuery** data into **Spark data frames** for advanced **ETL** capabilities.
* Built reports for monitoring data loads into **GCP** and drive reliability at the site level.
* Participated in daily stand-ups, bi-weekly scrums and PI panning.

**Environment:** Hadoop , GCP, BigQuery, Big Table, Spark , Sqoop , ETL, HDFS, Snowﬂake DW, Oracle Sql, MapReduce, Ka a and Agile process.

# Cyient- India Apr 2016- Jul 2018

**Data Engineer**

# Responsibilities:

* + Worked as **Data Engineer** to review business requirement and compose source to target data mapping documents.
	+ Conducted technical orientation sessions using documentation and training materials.
	+ Gathered the business requirements from the Business Partners and Subject Matter Experts.
	+ Served as technical expert guiding choices to implement analytical and reporting solutions for client.
	+ Worked closely with the **business, other architecture team members** and global project teams to understand, document and design **data warehouse** processes and needs.
	+ Implemented Installation and conﬁguration of multi-node cluster on Cloud using **Amazon Web Services (AWS)** on

**EC2**.

* + Developed reconciliation process to make sure **elastic search** index document count match to source records.
	+ Maintained **Tableau** functional reports based on user requirements.
	+ Created action ﬁlters, parameters, and calculated sets for preparing dashboards and worksheets in **Tableau.**
	+ Used **Agile (SCRUM)** methodologies for Software Development.
	+ Developed data pipelines to consume data from **Enterprise Data Lake (MapR Hadoop distribution - Hive tables/HDFS)** for analytics solution.
	+ Created **Hive** External tables to stage data and then move the data from Staging to main tables.
	+ Wrote **complex Hive queries** to extract data from heterogeneous sources **(Data Lake)** and persist the data into **HDFS**.
	+ Implemented the **Big Data** solution using **Hadoop, hive** and **Informatica** to pull/load the data into the **HDFS** system.
	+ Developed incremental and complete load **Python** processes to ingest data into **Elastic Search** from **Hive.**
	+ Pulled the data from data lake **(HDFS)** and massaging the data with various **RDD** transformations.
	+ Created **Oozie workﬂow** and Coordinator jobs to kick oﬀ the jobs on time for data availability.
	+ Developed Rest services to write data into **Elastic Search** index using **Python** Flask speciﬁcations
	+ Developed complete end to end **Big-data** processing in **Hadoop** eco system.
	+ Used **AWS Cloud** with Infrastructure Provisioning / Conﬁguration.
	+ Used **Hive** to analyze the partitioned and **bucketed** data and compute various metrics for reporting on the dashboard.
	+ Created dashboards for analyzing **POS** data using **Tableau**.
	+ Developed **Tableau visualizations** and dashboards using **Tableau Desktop**.
	+ Involved in **PL/SQL** query optimization to reduce the overall run time of stored procedures.
	+ Used **Hive** to analyze the partitioned and **bucketed** data and compute various metrics for reporting on the dashboard.
	+ Continuously tuned **Hive UDF's** for faster queries by employing partitioning and **bucketing**.
	+ Implemented partitioning, dynamic partitions and buckets in **Hive**.
	+ Deployed **RMAN** to automate backup and maintaining scripts in recovery catalog.
	+ Worked on **QA** the data and adding **Data sources, snapshot, caching** to the **report.**

**Environment:** AWS, Python, Agile, Hive, Oracle 12c, Tableau, HDFS, PL/SQL, Sqoop , Flume

# Innvonix Tech Solutions - India June 2012- Mar 2016

**Data Anlyst Responsibilities:**

* Eﬀectively led client projects. These projects contained a heavy Python, SQL, Tableau and data modelling.
* Performed data merging, cleaning, and quality control procedures by programming data object rules into a database management system.
* Created detailed reports for management.
* Reported daily on returned survey data and thoroughly communicated survey progress statistics, data issues, and their resolution.
* Involved in Data analysis and quality check.
* Extracted data from source ﬁles, transformed and loaded to generate CSV data ﬁles with Python programming and SQL queries.
* Stored and retrieved data from data-warehouses.
* Created the source to target mapping spreadsheet detailing the source, target data structure and transformation rule around it.
* Wrote Python scripts to parse ﬁles and load the data in database, used Python to extract weekly information from the ﬁles, Developed Python scripts to clean the raw data.
* Worked extensively with Tableau Business Intelligence tool to develop various dashboards.
* Worked on datasets of various ﬁle types including HTML, Excel, PDF, Word and its conversions.
* Analysed data from company databases to drive optimization and improvement of product development, marketing techniques and business strategies
* Performed Database and ETL development per new requirements as well as actively involved in improving overall system performance by optimizing slow running/resource intensive queries.
* Developed data mapping documentation to establish relationships between source and target tables including transformation processes using SQL.
* Participated in data modelling discussion and provided inputs on both logical and physical data modelling.
* Reviewed the Performance Test results to ensure all the test results meet requirement needs.
* Created master Data workbook which represents the ETL requirements such as mapping rules, physical Data element structure and their description.

**Environment**: Oracle 10g, UNIX Shell Scripts, MS Excel, MS Power Point, Python , SQL.