**Chopang Gurung K**

**Senior Data Engineer**

**Email** = cgk897@gmail.com

**Mobile** = +1 3014786684

**Professional Summary:**

* Holding **12 years** of experience in the field of **Data Engineering**.
* Database development including **Design architecture, development, system integration** and **infrastructure readiness**.
* Extensive experience in managing & delivering projects, conducting business analysis to determine requirements and planning for systems implementation at the enterprise scale an effective communicator with honed relationship, leadership, planning, coordination and analytical skills with ability to view the larger picture, taking business - critical decisions from the front to attain business/ operational excellence.
* Handling end-to-end implementing solutions on enterprise level from requirement analysis to **system study, designing, coding testing, de-bugging, documentation** and implementation extended experience in product-based companies.
* Design and implement Database architecture solutions for various application needs.
* Analyse multiple sources of structured and unstructured data to propose and design data architecture solutions for scalability, high availability, fault tolerance, and elasticity.
* Hands on experience in developing **ETL/ ELT** pipelines using **Python** & **Spark.**
* Experience in developing **Rest APIs** and **modelling data from API sources.**
* Develop conceptual, logical and physical design for various data types and large volumes.
* Solid expertise in **RDBMS** and Dimensional Modelling, Designing of **STAR, SNOW FLAKE Schemas.**
* Performance tuning **ETL Packages**, Stored procedures and **Adhoc queries** by applying standard practices, indexing strategies along with good understanding of tool & server architecture.
* Developed numerous reports and complex dashboards in **Power BI, Tableau** etc.,
* Architect, design and implement high performance large volume data integration processes, database, storage, and other back-end services.
* Hands-on expertise in developing Databases, Complex stored procedures, database objects (Tables, Views, Triggers etc.,), **ETL packages & OLAP Cubes** (Tabular and Multi-Dimensional) using Microsoft Intelligence BI Tools **(SSIS, SSAS)** Work closely with Business, at a technical and user level, to design and produce solutions.
* Work closely with the product management and development teams to rapidly translate the understanding of customer data and requirements to product and solutions.
* Participate in Rapid Application Development and Agile processes to deliver new Data warehouse services and components.
* Set architectural vision and direction across a matrix of teams.
* Design, development & maintenance of the **ETL** frameworks, Business Intelligence Solutions.
* Responsible for all proactive /reactive response for Proposals across the different line of Businesses.
* Providing Inputs for Pre sales team on different proposals.
* Responsible of meeting practice objectives such as Recruitment Support & Team Interface, Handling Proof of concepts on technology space, Building competencies.
* Sharing best practices, reviewing Projects from other Accounts.

**Technical Skills:**

**Operating Systems:** Windows, Mac OS and Linux CentOS, Debian, Ubuntu

**Programming Languages:** Python, R, C, C++

**Web Technologies:** HTML/HTML5, CSS/CSS3, XML, jQuery, JSON, Bootstrap, Angular JS

**Python Libraries/Packages:** NumPy, SciPy, Boto, Pickle, PySide, PyTables, Data Frames, Pandas, Matplotlib, SQLAlchemy, HTTPLib2, Urllib2, Beautiful Soup, Py Query

**Statistical Analysis Skills:** A/B Testing, Time Series Analysis, Marko

**IDE:** PyCharm, PyScripter, Spyder, PyStudio, PyDev, IDLE, NetBeans, Sublime Text, Visual Code

**Machine Learning and Analytical Tools:** Supervised Learning (Linear Regression, Logistic Regression, Decision Tree, Random Forest, SVM, Classification), Unsupervised Learning (Clustering, KNN, Factor Analysis, PCA), Natural Language Processing, Google Analytics Fiddler, Tableau.

**Cloud Computing:** AWS, Azure, Rackspace, OpenStack

**AWS Services:** Amazon EC2, Amazon S3, Amazon Simple DB, Amazon MQ, Amazon ECS, Amazon Lambdas, Amazon Sage maker, Amazon RDS, Amazon Elastic Load Balancing, Elastic Search, Amazon SQS, AWS Identity and access management, AWS Cloud Watch, Amazon EBS and Amazon CloudFormation

**Databases/Servers:** MySQL, SQLite3, Cassandra, Redis, PostgreSQL, CouchDB, MongoDB, Teradata, Apache Web Server, NginX, Tomcat, JBoss, WebLogic

**ETL:** Informatica, Data Stage, SSIS.

**Web Services/ Protocols:** TCP/IP, UDP, FTP, HTTP/HTTPS, SOAP, Rest, Restful

**Miscellaneous:** Git, GitHub, SVN, CVS

**Build and CI tools:** Docker, Kubernetes, Maven, Gradle, Jenkins, Hudson, Bamboo

**SDLC/Testing Methodologies: Agile, Waterfall, Scrum, TDD**

**Professional Experience**

**Capital One, New York Nov 2022 – Present**

**Lead data Engineer**

**Description:** Developed and maintained Python scripts for data collection and report development, working with **Azure** for **Hadoop** application deployment. Implemented fraud detection rules through **PQL** and optimized data ingestion pipelines, reducing code complexity. Designed and implemented analytics solutions for web data, facilitating real-time reporting. Provided technical leadership, coordinated offshore teams, and ensured seamless integration of data solutions for business insights. Expertise in **Oracle, Hadoop, SQL Server, Python**, and various web technologies.

**Responsibilities**

* **Python Script** reviewing for the data collection, reports development scripts, altercation. Made many different changes in the functionality.
* Created data frames for data type testing, did normalization analysis on data supporting data science team.
* Worked in **Azure** environment for development & deployment of **Custom Hadoop** Applications.
* Wrote pre-processing queries in python for internal **spark jobs**.
* Worked on different kinds of manual rule writing which detected fraud through the internal web application using **PQL**.
* Loaded data from many different sources to the **PostgreSQL database (Greenplum)** which was parallel processed & integrated data science repository with the environment creating **CI/CD pipelines**.
* Wrote Table Schemas for **Greenplum Database**.
* Supported architectural designing with the architects technically and by designing in **Visio**.
* Performed real time integration and loading data from the **Azure data box**& mounting it onto fuse for bulk loads.
* Wrote key spaces scripts in **Cassandra** and monitored cluster to create replication onto the database which follows **gossip protocol**.
* Created complex **SSIS packages** building pipelines and **ETL** for bulk & daily data injection.
* Worked on **Hadoop HDInsight** to inject Data Science Scoring models to **Greenplum (Data Lake).**
* Created Sqoop jobs to inject TBs of data from **RDBMS multiple to Hadoop**.
* Support data ingestion, using shell and **pySpark scripts**, into a **Hadoop Database**.
* Support includes production support as well as creation of code to ingest data for new lines of business.
* Re-engineered data ingestion pipeline reducing the amount of code to manage.
* Developed complex shell scripts to support the deployment and implementation of updated **Hadoop/Hive tables** and application functionality.
* Developed complex **SQL scripts** to not only ingest customer data into the **Customer Hub**, but to make that data available to make that data available to end users by creating custom extracts.
* Developed **PL/SQL** solutions to resolve issues with data in Operational Data Store Created a library of reusable types, functions, and procedures provide for developers to be implemented into all new scripts to help ensure script reliability and improve data reliability.
* Helped develop the data ingestion infrastructure for business-facing tool to allow for the quick delivery of customized data Wrote complex queries and stored procedures using **PL/SQL for Oracle** as well **as T-SQL in SQL Server.**
* Developed prototypes using new or expanding on existing technologies.
* Design, develop and implement the processes, solution to scrape inventory and pricing statistics from competitor websites.
* Analyze the gathered requirements and document the technical requirements.
* Work with the project manager and put together a plan for the delivery on the execution side of the roadmap.
* Define the application’s architecture and detailed design as per the approved design patterns.
* Create detail design document and develop the required code.
* Co-ordinate, delegate and supervise the offshore development team.
* Design and implement the whole analytics data solution for the gathered data from various web analytics platforms and scrapping data stores catered to produce the required reports/dashboards every two hours.
* Participate in the system, integrated and user acceptance testing activities.
* Demo/showcase the products to the partner teams and act as a **SME** to make sure they understand the tasks to consume and utilize the data/products.

**Environment:** Oracle databases, Hadoop, Hive, Hive SQL, Spark, pySpark, Python, Unix Shell Scripting, PL/SQL, MS SQL Server, MS Access, MS Excel, MS SQL Reporting Services, Stored Procedures (T-SQL), Telerik ASP .NET Web Controls, Java, WebLogic, C#, Gosu, Eclipse, jQuery, jQuery Tools, JavaScript, Infragistics ASP .NET Web Controls, VB .NET, AJAX, Cascading Style Sheets (CSS), Photoshop, MS Visual Studio, Subversion Source Control, Red Gate SQL Developer Bundle, SSMS SQL Tools, NUnit, MS Word

**Kaiser Permanente, Oakland, CA June 2017 – Oct 2022**

**Senior Data Engineer**

**Description:** Designed and implemented a scalable big data solution leveraging Hadoop ecosystem tools such as **Hive, MongoDB, and Spark Streaming** for real-time data processing. Utilized **Agile methodologie**s for efficient project management and collaborated in daily scrum meetings. Engineered data pipelines with **Kafka and Spark Streaming** to ingest, process, and store data in Parquet format on **HDFS**. Implemented various data transformation techniques using **Pig and Hive scripts, Sqoop, and custom-built Java MapReduce jobs**. Integrated analytical capabilities into cloud architecture, supported by technologies **like Apache Nifi and Neo4J** for data cleansing, visualization, and insights generation.Top of Form

**Responsibilities:**

* As a Data Engineer worked on Hadoop eco-systems including **Hive, MongoDB, Zookeeper, Spark Streaming with MapR distribution**.
* Involved in Agile methodologies, daily scrum meetings, sprint planning.
* Used Different Spark Modules like **Spark core, Spark SQL, Spark Streaming, Spark Data sets and Data frames.**
* Worked on **MongoDB** by using **CRUD** (Create, Read, Update and Delete), **Indexing**, **Replication** and **Shading** features.
* Extract Real time feed using **Kafka and Spark Streaming** and convert it to **RDD** and process data in the form of Data Frame and save the data as **Parquet format in HDFS**.
* Scripts were written for distribution of query for performance test jobs in **Amazon Data Lake**.
* Created **Hive Tables**, loaded transactional data from **Teradata** using **Sqoop** and worked with highly unstructured and semi structured data of 2 Petabytes in size.
* Implemented **Apache Nifi flow** topologies to perform cleansing operations before moving data into **HDFS**.
* Designed **novel machine learning algorithms** for predictive modelling and produce data insights (Data Science).
* Implemented multiple **MapReduce Jobs** in java for data cleansing and pre-processing.
* Build Hadoop solutions for bigdata problems using **MR1 and MR2 in Yarn**.
* Handled importing of data from various data sources, performed transformations using **Hive, Pig, and loaded data into HDFS.**
* Involved in identifying job dependencies to design workflow for **Oozie & Yarn** resource management.
* Configured Oozieworkflow to run multiple **Hive and Pig** jobs which run independently with time and data availability.
* Imported and exported the analyzed data to the relational databases using **Sqoop** for visualization and to generate reports for the BI team.
* Support Cloud Strategy team to integrate analytical capabilities into an overall cloud architecture and business case development.
* Designed and implemented incremental ingestion strategy for various sources.
* Expertise in writing **Pig Latin, Hive Scripts** to extended their functionality using User Defined
* Implemented Sqoop to bring data from various source databases like **Netezza, DB2, SQL Server**.
* Implemented incremental data load strategy from full load to change data capture approach.
* Created schema comparator program to identify the schema changes during incremental load.
* Created Hive tables to load **json files** and **avro data sets**.
* Identifying the partition column for data sets for **Hive tables**.
* Implemented pig program to load data from landing zone to Hive tables.
* Created data cleansing **UDF** in Pig to clean the data and add metadata columns.
* Created scripts to load data to Hive tables via **dynamic partition**.
* Implemented program to merge small lookup files to load as **json Hive tables**.
* Developed **POC** to load data from external websites **(NOAA)** to ingest data via **Apache NIFI**.
* Developed POC to visualize Hive data using **Neo4J.**

**Environment:** HIPAA, Hadoop, Hive, Sqoop, Pig, java, Django, Flask, XML, MySQL, MS SQL Server, Linux, Shell Scripting, Mongo dB, SQL, Python 3.3, Django, HTML5/CSS, Cassandra, JavaScript, PyCharm, GIT, Linux, Shell Scripting, RESTful, Docker, Jenkins, JIRA, jQuery, MySQL, Bootstrap, HTML5, CSS, AWS, EC2, S3.

**The Home Depot, Atlanta, GA Oct 2014 – May 2017**

**Senior Data Engineer**

**Description:** Implement a robust Continuous Integration and Continuous Delivery process using **GitLab** and **Python/Shell scripting** to automate routine tasks. Utilize **Django** framework and **Airflow** for efficient workflow management. Develop **AWS Lambda** functions for processing **nested JSON files**, while constructing **AWS data pipelines** with a focus on scalability, fault tolerance, and cost efficiency. Employ AWS services such as **Redshift, Data Lake, and DynamoDB** for data management and analysis, alongside implementing Restful and **SOAP APIs** for seamless integration.

**Responsibilities:**

* Implement **Continuous Integration and Continuous Delivery** process using **GitLa**b along with **Python and Shell scripts** to automate routine jobs, which includes synchronize installers, configuration modules, packages and requirements for the applications.
* Hands on **Django frame work** using **PyCharm**, hands on Airflow workflow management.
* Written **AWS Lambda** code in Python for nested Json files, converting, comparing, sorting etc.
* Construct the AWS data pipelines using **VPC, EC2, S3, Auto Scaling Groups (ASG), EBS, Snowflake, IAM, CloudFormation, Route 53, CloudWatch, CloudFront, CloudTrail.**
* Create and configure elastic load balancers and auto scaling groups to distribute the traffic and to have a cost efficient, fault tolerant and highly available environment.
* Manage metadata alongside the data for visibility of where data came from, its linage to ensure and quickly and efficiently finding data for customer projects using **AWS Data Lake** and its complex functions like **AWS Lambda, AWS Glue.**
* Develop the software systems, using scientific analysis, complex and optimal algorithms like sessionization for generating the statistics into web site analyzer.
* Develop Restful and **soap API’s** using swagger and perform Mobile app and customer product details app testing using **Postman**.
* Hands on with **Redshift Database** (**ETL data pipe lines** from **AWS Aurora** - **MySQL** Engine to Redshift).
* Distkey, Sort key, groups on Primary columns for better query performance and tuning.
* Assigning user level/ group level permissions on Redshift schema for security reasons
* Restoring **PostgreSQL** into **Redshift** (except certain data types).
* **AWS DMS** replication instance between different databases.
* Design and construct of **AWS Data pipelines** using various resources in AWS including **AWS API Gateway** to receives response from AWS lambda and retrieve data from snowflake using lambda function and convert the response into Json format using Database as **Snow Flake, DynamoDB, AWS Lambda function and AWS S3.**
* Develop and implement complex databases and data marts for current Production state for both traditional **RDBMS** and **Hadoop Eco system** with the legacy applications and designing the solutions to incorporate new processes & implementation into existing environment.
* Build **ETL pipeline** end to end from **AWS S3 to Key**, Value store **DynamoDB**, and **Snowflake Datawarehous**e for analytical queries and specifically for cloud data.
* Convert into different Data formats for user/business requirements by streaming data pipeline from various sources Snowflake and unstructured data, **Dynamo-db.**
* Perform document modifications and enhancements made to the applications, systems and databases as required by the project.
* Document all the changes implemented across all systems and components using Confluence and Atlassian Jira. Documentation includes technical changes, Infrastructure changes, and Business Process changes. Post Release documentation would also include Known Issues from Production Implementation and Deferred defects.
* Perform Informatica Cloud Services, Informatica Power Centre Administration ETL strategies and ETL Informatica mapping. Setting up of Secure Agent and connect different applications and its Data Connectors for processing the different kinds of data including unstructured (logs, click streams, Shares, likes, topics etc..), semi structured (XML, JSON) and structured like **RDBMS**.

**Environment:** Hadoop, Hive, Zookeeper, MapR, Agile, Teradata, MapReduce, Yarn, HDFS, Sqoop, Kafka, Apache Nifi, MongoDB, DTCC, Pig, HBase, Python, AWS, Scala, Spark, Oozie, Yarn, Cassandra

**Caterpillar, Peoria, IL Feb 2012 – Sep 2014**

**Data Engineer**

**Description:** Designed and implemented end-to-end data pipelines on **GCP** and **AW**S. Proficient in **ETL** processes, employing tools like **Apache Airflow, Docker, and Kubernetes**. Skilled in machine learning and data visualization, with expertise in **Python, Spark, and Big Query** for scalable data processing and analysis.

**Responsibilities:**

* Experience in building and architecting multiple Data pipelines, end to end **ETL** and ELT process for Data ingestion and transformation in **GCP.**
* Strong understanding of AWS components such as **EC2 and S3.**
* Implemented a Continuous Delivery pipeline with **Docker and Git Hub.**
* Worked with g-cloud function with Python to load Data in to big query for on arrival csv files in **GCS bucket**.
* Process and load bound and unbound Data from **Google pub**/sub topic to big query using cloud Dataflow with **Python**.
* Devised simple and complex **SQL scripts** to check and validate Dataflow in various applications.
* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through **Python**.
* Developed and deployed data pipeline in cloud such as **AWS and GCP**.
* Performed data engineering functions: data extract, transformation, loading, and integration in support of enterprise data infrastructures - data warehouse, operational data stores and master data management.
* Responsible for data services and data movement infrastructures good experience with **ETL** concepts, building **ETL solutions** and **Data modelling.**
* Hands on experience on architecting the ETL transformation layers and writing spark jobs to do the processing.
* Gather and process raw data at **scale** (including writing scripts, **web scraping, calling APIs, write SQL** **queries**, writing applications).
* Experience in fact dimensional modelling (**Star schema, Snowflake schema**), transactional modelling and **SCD** (Slowly changing dimension).
* Devised **PL/SQL Stored Procedures**, Functions, Triggers, Views and packages. Made use of Indexing, Aggregation and Materialized views to optimize query performance.
* Developed logistic regression models (**Python**) to predict subscription response rate based on customers variables like past transactions, response to prior mailings, promotions, demographics, interests, and hobbies, etc.
* Develop near real time data pipeline using **spark**.
* Process and load bound and unbound Data from **Google pub/sub** topic to big query using cloud Dataflow with **Python**.
* Hands of experience in **GCP**, **Big Query, GCS bucket, G - cloud function, cloud dataflow, Pub/suB cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver.**
* Implemented Apache Airflow for authoring, scheduling and monitoring Data Pipelines
* Proficient in Machine Learning techniques (Decision Trees, Linear/Logistic Regressors) and Statistical Modelling.
* Worked on confluence and Jira skilled in data visualization like **Matplotlib** and **seaborn librar**y.
* Hands on experience with big data tools like **Hadoop, Spark, Hive**.
* Experience implementing machine learning back-end pipeline with **Pandas, NumPy**.

**Environment:** Big query, Bucket, G-Cloud Function, Apache Beam, Cloud Dataflow, Cloud Shell, Gsutil, Docker, Kubernetes, AWS, Apache Airflow, Python, Pandas, Matplotlib, seaborn library, text mining, NumPy, Scikit-learn, Heat maps, Bar charts, Line charts, ETL workflows, linear regression, multivariate regression, Python, Scala, Spark

**Education:**

Bachelors in Computer Science,

University of Maryland Eastern Shore - 2011