**Om Prakash Karmacharya**

**Big Data Engineer**

[ompkarmacharya@gmail.com](mailto:ompkarmacharya@gmail.com) | (573) 222-0175

<https://www.linkedin.com/in/om-karmacharya/>

**Professional Summary**

* Overall, 5+ years of professional experience in Data Analytics, Big data, knowledge of Hadoop Framework, Hadoop, and parallel processing implementation.
* Hands on experience on **Hadoop** /**Big** **Data** related technology experience in Storage, Querying, Processing, and analysis of data.
* Expertise in **Data mining**with large datasets of Structured and Unstructured Data, Data Acquisition, Data Validation, Predictive modeling, Data Visualization.
* Experience in the development of **Big Data** projects using **Hadoop**, **Hive**, **HDP**, **Pig**, and **MapReduce** open-source tools.
* Hands-on experience in installing, configuring, and using Hadoop components like **Hadoop** **MapReduce**, **HDFS**, **Hive**, **Sqoop**, **Pig**, **Zookeeper,** and **Flume**
* Development of **spark**-**based** application to load streaming data with low latency, using **Kafka**.
* Experience in developing a data pipeline through **Kafka**-**Spark** **API**.
* Experience in importing and exporting the data using **Sqoop** from **HDFS** to Relational Database systems and vice-versa and load into **Hive** **tables**, which are partitioned
* Hands on **Spark** **MLlib** utilities
* Experience working with **NoSQL** databases such as **HBase** and **Cassandra** to store structured, semi-structured, and unstructured data.
* Experience with **Flume** to load the log data from multiple sources directly into **HDFS.**
* Experience with **Cloudera**, **Hortonworks** and **MapR** distributions.
* Good Knowledge in Amazon Web Service (**AWS**) concepts like **EMR** and **EC2** web services successfully loaded files to **HDFS** from **Oracle**, **SQL** **Server** and **Teradata** using **Sqoop**.
* Development of **Spark**-based application to load streaming data with low latency, using **Kafka** and **Spark** programming
* Hands-on experience in **cleansing, transformation, and visualization** of the data using Python.
* Good experience with python and its libraries like **NumPy, Pandas, Matplotlib, Seaborn, NLTK, Sci-Kit learn, and SciPy.**
* Possess good experience in Pattern Mining, Outliner Detection, Clustering for Descriptive Analysis Problems.
* Experience in **Data Migration** from one database source to other.
* Experienced in agile/iterative development process to drive timely and impactful data science deliverables.

**Career Competencies:**

|  |  |
| --- | --- |
| **Languages** | Python 2.7.x and Python 3.x, SQL, PL/SQL, Shell Scripting, Storm 1.0, JSP, Servlets, Scala, Python, Java, R, JavaScript |
| **Big Data Technologies:** | Hadoop 3.0, HDFS, Map Reduce, HBase 1.4, Apache Pig, Hive 2.3, Sqoop 1.4, Apache Impala 2.1, Oozie 4.3, Yarn, Apache Flume 1.8, Kafka 1.1, Zookeeper |
| **Databases**: | SQL, Spark SQL, My SQL, MS Access, HDFS, HBase, Oracle 12c/11g |
| **Project Execution Methodologies:** | Kimball data warehousing methodology, Agile Scrum Methodology, CRISP-DM |
| **Regression**: | Linear Regression, Ridge Regression, Polynomial Regression, Lasso Regression, Elastic Net |
| **Clustering**: | k-Means, Hierarchical Clustering, Latent Dirichlet Allocation (LDA) |
| **Cloud Platform:** | Amazon Web Services, Microsoft Azure |
| **Version Control:** | GIT, SVN, CVS |

**Education Details**

**Bachelors in computer science (Evaluated by WES)**

**Kathmandu University**

**Career Experiences:**

**Comcast,** **Richmond, VA** **Jan 2021 - Present**

**Big Data Engineer**

* Remained highly involved throughout all phases of the project i.e., **Project Planning and Problem Definition**, **Data Engineering,** **Data Collection, and Analysis**, **Model Development and Selection** **Evaluation, and Deployment**
* Develop **Spark** **core**, **Spark** **SQL**/scripts using **Python** and **Scala** for faster data processing and use HBase to load the data.
* Intake happens through Sqoop, and Ingestion happens through **Map** **Reduce**, **HBASE**.
* Working on Implementing incremental logic to import data using **Sqoop** from **SQL** **Server** to **HDFS**.
* Monitored and implemented code to reprocess the failure message in **Kafka** using offset id.
* Building an **ETL** pipeline to stage data to **ADLS** for Snowflake ingestion and transformation.
* Brainstorming ideas on how to update to results to the external table for data consumption by the **ML** models.
* Developed **Hive** scripts for analyzing data.
* Monitored and updated **Oozie** workflows and different shell and **Spark** actions that run several daily processes.
* Extensively worked on Spark Streaming and **Apache** **Kafka** to fetch live stream data.
* Installed Kafka manager for consumer lags and for monitoring **Kafka** Metrics also this has been used for adding topics, Partitions.
* Divided application into two parts i.e., training and inference. Used **Sagemaker**, **AWS Batch** for training section, and **AWS Lambda** for inference part.
* Packaged code and required resources into a single deployable package using **Shell-Script**, **Docker Containers**, **CloudFormation, Python Scripts**
* Participated in multi team’s effort to build an energy-specific **language model** like BERT from google using **Attention models.**
* Communicated process, results, and possibilities to key stakeholders like CIO, Product Owner, Business users using a well-articulated PowerPoint presentation.

**Data Engineer**

**Deerwalk Inc., Boston, MA**  **Mar 2017 – Dec 2020**

* Developed Spark code and **Spark-SQL/Streaming** for faster testing and processing of data.
* Utilized experience and expertise in **Apache** **Spark** ecosystem to build a **lambda** architecture that could handle massive volumes of data in real-time
* Worked on a huge volume of Data in **Hadoop** Cluster and Splunk to build an event clustering model to find anomalous logs from 100+ applications.
* Implement Data Exploration to analyse patterns and to select features using **SparkSQL** and other **PySpark** libraries.
* Created external **HIVE tables** for analytical querying on the data present in HDFS
* Querying **SQL database** for customer production issue resolutions.
* Write the **Lambda** **function** to read the ctrl files from **S3** whenever its triggered and compare the missing records from **Database**
* Created import and scrub script using **MS** **SQL** for data import from **AWS S3** and cleansing as an **ETL** process.
* Matching member information across data sources, standardization (e.g., standardization of relationship codes/coverage, **LOA** [Level of aggregation]), and preparation of crosswalks files for intermediate process. Generation of reports for Monthly Member Month based on client **LOAs**, distributions based on claim factors (provider, claim number, amount, services, and paid dates). Analysis of **Healthcare** **KPIs** & various Metrics reports mainly from Medical and **Rx** data types relating with data types.
* Building up **SAS/Macros**, write metadata scripts in SQL server and setup **ETL** jobs in **SAS** to handle and process raw data files of various formats.
* Develop, validate, and implement **SAS** programs and produce derived datasets for analysis and generating and documenting tables, **DQR** to study reports as well as share information with clients.
* Developed and maintained data metrics, data sets, reports (**Data** **Standardization** **Document** (**DSD**), **Data** **Quality** **Report**, and **Import** **Quality** **report**), dashboards, to inform decision making and drive continual improvements.
* Prepared custom **analytical**/**quality** **reports** as per business needs frequently (daily, weekly, bi-monthly, monthly).
* Developed data standard document with a set of rules to convert raw data into standard format data.

**References Available Upon Request**