Name: Hitha
Ph. No: 647-849-2058
Mail id: hitha.developer@gmail.com

SUMMARY:

- **5+ years** of experience in computer science, especially **Big Data**, **Scala.**
- Worked in multiple fields including music, finance, insurance, and telecommunication
- Worked under **Big Data** eco-systems include **Google Cloud Platform**, **Cloudera CDH** and **Hortonworks HDP**
- Extensive knowledge of **Big Data** with **Hadoop, YARN, HDFS, Google Cloud Storage, MapReduce**, **Spark**, **Beam**, **Pub/Sub**, **Kafka**, **BigTable**, **HBase, BigQuery** and **Hive**
- Experienced in languages such as **Scala, SQL**
- Experienced with **Real-time** data processing mechanism such as **Google Cloud Pub/Sub**, **Apache Kafka** and **Spark Streaming**
- Expertise in **Apache Beam Scio** (**Scala**) **API**, **Spark Scala API, MapReduce** with **Scalding** (**Scala**) to create pipelines for data **ETL** and model building
- Developed data models that powers content grouping, ranking, mapping and predicting
- Involved in writing **Query** on **BigQuery** and **Hive** with UDF for data analyzing and evaluation
- Good knowledge of working with **Docker** containers
- Experienced in dependencies management and workflow scheduling in **Big Data** ecosystem
- Conducted data transformation with data formats like **Avro**, **Parquet** and **Sequence File**
- Adept at using **Sqoop** to migrate data between **RDBMS**, **NoSQL** data bases and **HDFS**
- Worked with **NoSQL** databases including **BigTable**, **HBase**, **Cassandra** and **MongoDB**
- Worked with **RDBMS** including **MySQL**, **Oracle SQL**
- Involved in developing **Machine Learning** algorithms including **Linear Regression**, **Logistic Regression**, **K-Means**, **Decision Trees**
- Worked with Machine Learning libraries including **NLTK**, **Scikit-learn**, **SciPy**
- Worked with **Windows** & **Linux** operating systems for development
- Strong knowledge of **Linux/Unix** Shell Commands
- Good knowledge of **Unit Testing** with S**calaTest, JUnit, MRUnit** and **Pytest**
- Familiar with developing environments like **JIRA, Confluence** and **Agile/Scrum**
- Successfully worked under high pressure and completed projects with tight deadlines
- A self-motivated learner, challenger and good team player

TECHNICAL SKILLS:

Programming Language	**Scala, SQL, JavaScript**
Hadoop Ecosystem	**YARN, HDFS, HBase, Hive, Sqoop, Pig, Flume, Zookeeper, Oozie.**
Web Technologies	**HTML5, CSS3, JavaScript, jQuery, AngularJS, Angular 7, 8, React JS, TypeScript, Node JS,**
Web Development	**Apollo, Hibernate, Spring, SOAP, REST**
Database Server	**MySQL, Oracle, MongoDB, HBase, Cassandra**
IDE	**Eclipse, IntelliJ, Visual Studio**
Unit Testing	**Scala Test, JUnit, MR Unit, Pytest**
Data Analysis & Visualization	**Scala, NLTK, Scikit-learn, \ ggplot, matplotlib,**
Data Pipeline Development	**Spark, Scio, Scalding, Crunch, MapReduce\ Pub/Sub, Kafka**
Google Cloud Platform	**Google Cloud Storage, Big Table, Big Query, Data Flow, Data proc, Pub/ Sub, Apache Beam.**
Building Tools	**Ant, Maven**

PROFESSIONAL EXPERIENCE:

Client: Genpact, Toronto, Ontario **Jun'19 – Till Date**
Big Data Engineer

Genpact Headstrong Capital Markets is a global consulting and IT services company with a specialized focus in capital markets. With more than 20 years of experience consulting with 9 of the world's top 10 investment banks, we are the world's leading technology services provider for the financial services and securities industries. This experience enables us to help transform capital markets business operations using a unique combination of our proprietary SEPSM performance benchmarks, process focus and technology expertise.

Responsibilities:

- Worked on **Google Cloud Platform** (**GCP**) with **Agile** development cycle
- Developed pipelines with **Scio** (**Scala**) **API** of **Apache Beam** for data **ETL** and model building
- Improved and modified **MapReduce Scalding**(**Scala**) and data pipelines with models for music content grouping/ranking and third party catalog mapping
- Designed **Google BigTable** schema with **Turtle** to store third party metadata
- Developed real-time models with **Scala** and **Pub/Sub** to consolidate third party metadata with Spotify entities in **BigTable** based system to support in-client features
- Migrated data pipelines from on premise **Hadoop** system onto **GCP**, dockerized and configured **Spark** jobs to run with **Cloud Dataproc**
- Modified **REST** services for transcoding and ingesting data to **Google Cloud Storage**
- Worked with data scientists to explore log data and extract content to power search model

- Configured with **Styx** to manage **Docker** container executions and batch job scheduling
- Written ad-hoc **SQL** on **Google Big Query** for analyzing and evaluating large datasets
- Performed unit testing with **ScioTest**, **ScalaTest** and **JUnit**
- Worked on maintaining system with multiple projects that powers more than **300** daily partitioned data storage endpoints
- Involved in designing products, evaluating work scopes and defining testing metrics
- Provided technical supports for other teams such as music editors, data scientists, downstream data consumers and teams responsible for in-client features
- Involved in building and deploying Apache **MAVEN** scripts, debugging through logging frameworks like Log4j, automated build tool with Jenkins.
- Apache **Maven** is used as Build tool to automate the build process for the entire application.
- Used **Git** for version control, **Jenkins** for continuous integration and **JIRA** for project tracking

Environment:
Google Cloud Platform, Google Cloud Storage, BigTable, BigQuery, Pub/Sub, Hadoop, HDFS, Cassandra, Docker, Luigi, Styx, Apache Beam, Scio, Spark, MapReduce, Scalding, Crunch, Scala, SQL, Maven

Client: IBM, markham, Ontario **May'18 - Jun'19**
Big Data Engineer

The project was to develop the application called Customer first (CF), where we will onboard third-party agents Like MOSS, TPAR, and GEICO to get access to bind the Quotes. In the CF application we develop CFA (Auto) and CFP (Property) quotes with miscellaneous functions access to the third-party agent. The prefill data is accessed to the agents selling the quotes to the customers will be controlled by CF application.

The Big Data and Analytics platform of aims at supporting the data science team to build & test models for detecting insurance fraud. The system takes structured and unstructured data from claims databases and handwritten adjuster notes to identify potential fraud.

Responsibilities:
- Worked on **Hortonworks Data Platform 2.x** with **Agile** methodology
- Designed and built **Hive** databases with partitioned and bucketed tables
- Extracted data from **MongoDB** through MongoDB Connector for **Hadoop**
- Used **Sqoop** to transfer data from **RDBMS** to **HDFS**
- Worked with multiple data formats (**Avro, Parquet, CSV, JSON**)
- Wrote customized **Hive UDFs, HiveQL** for data retrieval and analyzing
- Worked with **Flume** to capture web server log data
- Developed **PIG Latin** scripts to transform data and load into **HDFS**
- Implemented predictive and statistical model with **Hadoop MapReduce**
- Performed unit testing using **Pytest**, **JUnit** and **MRUnit**
- Used **Git** for version control and **JIRA** for project tracking
- Apache **Maven** is used as Build tool to automate the build process for the entire application.

Environment: Hortonworks HDP, Hadoop, HDFS, Hive, Pig, Flume, Sqoop, Oracle, MySQL, MongoDB, HiveQL, Maven.

Client: MAQ Software - Hyderabad, India **Aug'15 - Mar'18**
Big Data Engineer

It is an IT Outsourcing company with expertise on mobility, enterprise applications, data modelling, business logic, UI/UE, E-Commerce has come a long way from being a start-up to becoming a 100% Export-Oriented Offshore Software Development Centre.

Responsibilities:

- Worked on **CDH 5.x** with **Agile** development cycle
- Designed **HBase** schema for the ingestion of streaming time series data
- Developed real-time data pipelines with **Kafka** to receive data from multi-sources
- Configured **Spark Streaming** with **Kafka** to clean, aggregate real-time data, then store processed data into **HBase**
- Wrote **Spark** and **Spark SQL** in **Scala** for data **ETL** and model building, also changed pipelines from **MapReduce** job to **Spark**
- Developed time series data analysis models with **PySpark**
- Used **Sqoop** to move data between **Oracle** and **HBase**
- Integrated **HBase** with **Hive**, and wrote **HiveQL** for data analysis and updates
- Transferred data from **HDFS** and created visualization for report
- Deployed workflows in **Oozie** for workflow scheduling and executions
- Performed unit testing with **ScalaTest, Scala Check, JUnit** and **Pytest**
- Used **Git** for version control, **JIRA** for project tracking and **Jenkins** for continuous integration

Environment:
Cloudera CDH, AWS, Hadoop, HDFS, HBase, Hive, Oracle, Sqoop, Oozie, Kafka, Spark, Spark Streaming, MapReduce, Scala, Ant.