**Siddhartha Pande**

|  |  |
| --- | --- |
|  | Email: siddharthapande88@gmail.com |
| [linkedin.com/in/siddhartha-pande](https://linkedin.com/in/siddhartha-pande/) | Mobile:+1-408-203-9739 |

# Summary

Results-driven Data Engineer with over 12 years of experience in big data processing, database management, and data visualization. Proficient in Hadoop, Apache Spark, SQL, and Tableau. Known for exceptional productivity, efficient task execution, and strong problem-solving skills. Proven ability to develop innovative data solutions in diverse environments.

# Education

## Institute of Engineering and Management July 2007 – June 2011

*Bachelor of Technology in Computer Science and Engineering Kolkata, India*

**Technical Skills**

* **On-premises Datawarehouse**: Hadoop, Hive, HDFS, Spark, PySpark, Teradata, Oracle 11g, Vertica
* **Cloud Datawarehouse:** Snowflake, Singlestore
* **Cloud**: GCP
* **Languages**: SQL, Python, Unix Shell Scripting, PL/SQL
* **Tools**: GIT, Jupyter Notebooks, Docker, Vessel
* **BI Tools:** Tableau, SAP BO
* **Scheduler:** Airflow, Automic/UC4, Autosys
* **RM Tools:** Kubernetes
* **Data Modeling Tool:** Erwin
* **SDLC Methodologies**: Agile, Waterfall

# Experience

|  |  |
| --- | --- |
| **Employer: TCSL, Client: Apple Inc** | **May 2021 – Current** |
| *Senior Data Engineer* | *Cupertino, CA* |

* Designed robust ETL pipelines to seamlessly integrate KPIs from various sources, including relational databases, file-based systems, and APIs, into user interface dashboards.
* Performed cost and rule-based analysis on cloud systems (Snowflake, SingleStore) to optimize computation and storage expenses, **resulting in a 40% reduction** in the organization’s application costs.
* Redesigned and implemented the application to migrate ETL processes from an on-premises system (Teradata) to Snowflake on AWS cloud.
* Implemented a solution to identify significant KPIs and automate the detection of data discrepancies within client’s business model.
* Enhanced code to harness the power of JDBC parallel processing, AQE in a spark environment.
* Developed **ETL pipeline to extract data** from relational database through Spark JDBC connectivity.
* Employed Spark for efficient data transformation: cleaning, aggregation, and normalization. Loaded transformed data into Snowflake using the Spark Snowflake connector for improved querying and reporting.
* Designed and implemented a vacuum Utility for cleaning up data files and log files associated with Delta tables using PySpark.
* Rewrote python loading logic to **leverage asynchronous processing for faster data loading.**
* Implemented Dynamic query generation in python script for quarterly data computation.
* Developed and deployed python docker image using linux base in Kubernetes pod to leverage isolation and portability

## Employer: Flexton, Client: Walmart Jul 2018 – Feb 2021

*Data Engineer Cupertino, CA*

* Conducted analysis, aggregation, and reporting of clickstream data, user session data, and product attribute data for Walmart and its acquired tenants. Ensured a seamless flow of high-quality data through end-to-end pipeline process.
* Optimized hive table performance by converting the file format to ORC, enabling improved compression, ACID property usage, and efficient schema evaluation for handling transactional data.
* Parsed raw data in Hive using regular expressions and Python-based UDF.
* Extracted data from multiple sources to analyze and identify customer behavior insights.
* Designed and implemented a **Scala-based Spark pipeline** to integrate externally sourced CSV files with HDFS, serving as the primary data source for Hive tables.
* Utilized performance tuning techniques, including **dynamic allocation, broadcast joins, and data repartitioning in Spark**, to optimize execution engine performance on large-scale datasets, resulting in improved throughput.
* Developed an automation framework using shell and Python scripts to validate ETL metadata prior to execution, preventing redundant data processing and performing sanity checks in the filesystem, which **reduced data loading issues by 30%.**
* Designed data model using dimensional modeling for improved efficiency and accessibility
* Developed scalable workflow utilizing spark and GCP bucket to support legacy report on hive-MR and analytics on **GCP data warehouse system bigquery using hive metastore.**

## Employer: Infosys Ltd, Client: Apple Inc Jan 2017 – Jun 2018

*Data Engineer Cupertino, CA*

* Analyzed user behavior on the Apple iTunes website and created interactive visualizations in Tableau to present key findings and statistics, as well as published and managed dashboard schedules in Tableau Server.
* Developed an advanced recommendation system using Kafka to deliver personalized updates on new albums based on users’ top three genre preferences.
* Achieved a **7% increase in user win-back subscriptions and generated 19% more sales in iTunes USA** store over the course of one quarter.
* Utilized Tableau to create interactive dashboards and worksheets by implementing action filters, parameters, and calculated sets.
* Integrated data from multiple sources to Teradata using ETL pipeline and implemented complex business logic in stored procedures through the creation of SQL objects, such as tables, views, and macros.
* Enhanced loading efficiency and transformed data effectively by implementing complex logic in Teradata procedure.
* Enhanced efficiency by employing partitioning and optimized indexing techniques.
* Designed and tested sales forecast provisioning system using Teradata PL/SQL stored procedures, functions, views, tables, macros etc.
* Conducted large-scale data import operations from diverse sources to Teradata RDBMS employing BTEQ

## Employer: Infosys Ltd, Client: Apple Inc Aug 2011 – Dec 2016

*Data Engineer Bhubaneswar, India*

* Developed a robust data transformation pipeline in Teradata for effective handling of clients’ billing, booking, and backlog data.
* Helped with UAT activities from initial stages to final implementation.
* Developed a Python tool for efficient trend analysis of booking data and automated email notifications to users, enabling them to identify and analyze key performance indicators from the BO tool.
* Managed development of forecast application, including requirement analysis and optimized design implementation.
* Assisted in building an automated framework for daily data reporting, facilitating seamless uploading of information in excel format to client’s repository.
* Enhanced efficiency of various Map reduces tasks and **reduced total execution time to just 45% of previous levels.**
* Developed Unix-based data validation framework for verifying accuracy between source and target tables.
* Utilized Fastload and Multiload utilities for copying production level data from UAT environment to create a real-time testing environment.

# Accomplishments

• 6.431x: Probability - The Science of Uncertainty and Data awarded by MIT MicroMaster Program