Surendra kapa

BIG DATA ARCHITECT/PYTHON DEVELOPER

Phone – (302) 252 – 8377 kapasurendra5@gmail.com

Professional Summary

* 8 years of IT experience in software Development and Analytical Solutions with hands-on experience in development and design of Python and related frameworks, full stack web.
* Over 6 years’ experience as Hadoop Developer with expertized knowledge of Hadoop framework, Hadoop Distributed file system and Parallel processing implementation, Hadoop Ecosystems HDFS, Map Reduce, Hive, Pig, Sqoop, Hue, Oozie, Impala and Spark.
* Excellent understanding / knowledge of Hadoop architecture and various components such as HDFS, Job Tracker, Task Tracker, Name Node, Data Node and Map Reduce programming paradigm.
* Experience and Expertise in ETL, Data analysis and designing data warehouse strategies.
* Upgraded Hadoop CDH to 5.x, and Hortonworks. Installed, Upgraded and Maintained Cloudera Hadoop-based software, Cloudera Clusters, Cloudera Navigator.
* Strong experience in analyzing large amounts of data sets writing PySpark scripts and Hive queries.
* Worked extensively with Data migration, Data cleansing, Data profiling, Data comparison, and ETL Processes features for data warehouses.
* Had good understanding on Snowflake, Redshift cloud technologies.
* Experience in using Python's packages like xlrd, numpy, pandas, matplotlib, Pyspark, scipy, scikit-learn and IDEs - PyCharm, Anaconda, Jupyter, IPython and visual studio.
* Experience in leading multiple efforts to build Hadoop platforms, maximizing business value by combining data science with big data.
* In depth knowledge on Amazon Web Services (AWS) cloud services like EC2, S3, EMR and VPC.
* Experienced with Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load Balancing, Auto Scaling, Cloud Front, CloudWatch, SNS, SES, SQS and other services of the AWS family.
* Experience in using Pentaho Data Integration (PDI) tool for data integration, OLAP analysis and ETL process
* Experience in writing Sub Queries, Stored Procedures, Triggers, Cursors, and Functions on MySQL and PostgreSQL database.
* Experience with Unit testing/ Test-driven Development (TDD), Load Testing.

Work Experience

**AMERICAN EXPRESS – New York, NY Mar 2019 to Present**

**Big data Developer/Data Engineer**

The American Express Company, also known as Amex, is an American multinational financial services corporation headquartered in Three World Financial Center in New York City.American Express offers world-class Charge and Credit Cards, Gift Cards, Rewards, Travel, Personal Savings, Business Services, Insurance and more. It's known for its top-notch customer service and some of the best rewards available. From ultra-elite business travel cards to everyday rewards, American Express has credit cards that are a good choice for many different types of cardholders

**Responsibilities:**

* Performed data analysis and build Hive queries to capture end to end campaign journey of a customers.
* Worked in optimizing the hive queries to reduce the run time.
* Developed framework using PySpark to execute hive queries and capture statistics.
* Worked in Data Ingestion project to build marketing analytics platform and designed tables in Hive, MYSQL using SQOOP.
* Written Pig Latin scripts to process data and loading data set to hive tables.
* Used Spark-SQL to load JSON data and create schema RDD and loaded it into Hive Tables and handled structured data using SparkSQL.
* Involved in HBASE set up and storing data into HBASE, which will be used for analysis.
* Used Spark for interactive queries, processing of streaming data and integration with popular NoSQL database for huge volume of data.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Scala and python.
* Closely worked with Kafka admin team to set up Kafka cluster on the QA environments.
* Implemented Kafka producer and consumer applications on Kafka cluster set up.
* Designed and implemented column family schemas of Hive and HBase within HDFS.
* Created scripts in Python which integrated with Amazon API to control instance operations.
* Build distributed, reliable, and scalable data pipelines to ingest and process data in real-time and dealing with fetching impression streams, transaction behaviors, clickstream data and other unstructured data.
* Worked with java Collection classes like List, set etc. and leveraged Java API to internal real time application.
* Used Pandas data frames for data cleansing and data scrubbing.
* Spent time in cleaning data as per business requirements using user defined functions.
* Being a developer, sourced large volumes of data from diverse data platforms into Hadoop platform

**Environment:** Hadoop, Hive, Pig, SparkSQL, python, PySpark, MySQL, Shell Scripting, Java, Kafka,pandas

JP MORGAN CHASE – Wilmington, DE May 2017 to Feb 2019

Data Architect

JP Morgan Chase is an American banking and financial services holding company headquartered in New York City. It is the largest bank in the US. It is a major provider of financial services, and according to Forbes magazine is the world's fourth largest public company based upon a composite ranking. The data in Webstats is accessible based on the Line of Business-like Investment Banking, Retail Banking, Home Finance, Auto Finance, Student Loans, Consumer Banking etc.

Responsibilities:

* Participated in developing marketing analytics platform and Hadoop data lake and took part in requirement discussions and BRD reviews.
* Worked in bringing multiple data feeds to the Greenplum as part of marketing analytics platform on business requirement.
* Persisted data using PostgreSQL as a relational database to store and retrieve.
* Developed External writable and readable tables in Greenplum to access data from Hadoop Data Lake.
* Developed shell scripts and python framework to perform end-to-end ETL flow for data.
* Worked with bulk load API to pull the customer opportunities data from salesforce using SOQL.
* Worked in designing tables in Hive, MYSQL using SQOOP and processing data like importing and exporting of databases to the HDFS.
* Hands on experience in ingesting data into Data Warehouse using various data loading techniques.
* Transferred and loaded datasets from Hive tables to Greenplum using GPHDFS PULL.
* Experienced working with salesforce.com sandbox environments.
* Hands on experience in running hive queries on spark shell.
* Used python framework to ingest data to the Hadoop hive tables and also performed Data validation check using Pyspark programs.
* Developed Pyspark code to read data from Hive, group the fields and generate XML files from spark data frame results.
* Experienced in building API to download the data from Google Cloud and DCM (Double Click campaign manager)
* Developed and updated the existing spark jobs written in Scala to leverage with business process and has strong hold on troubleshooting Spark issues.
* Automated Jobs using control M scheduler.
* Worked in creating S3 buckets also managing policies for S3 buckets and Utilized S3 bucket and Glacier for storage and backup on AWS.
* Experience to manage IAM users by creating new users, giving them a limited access as per needs, assign roles and policies to specific user.
* Experienced in working with multiple databases like MYSQL, ORACLE, TERADATA, SQL and SYBASE as part of ETL process.
* Developed new application to support data feature engineering concepts using C sharp with spark as backend.
* Experienced using spark to ingest data from Teradata to Hadoop as Hive tables.

**Environment:** Cloudera Hadoop Distribution (CDH5.6), HDFS, Hive, HUE, Impala, Sqoop, Python 2.7, MySQL, MS SQL Server, Linux, Shell Scripting, Teradata, oracle, PostgreSQL, Greenplum, Spark, C #.

EASTBAY/FOOTLOCKER – Wausau, WI July 2016 to March 2017

Sr Python Developer with Hadoop

Eastbay is a supplier of athletic footwear, apparel and sports equipment, which sells by direct mail. Since 1997, it has been the direct-to-mail division of Foot Locker, Inc.

Responsibilities:

* Participated to develop a data platform from scratch and took part in requirement gathering and analysis phase of the project in documenting the business requirements.
* Worked in designing tables in Hive, MYSQL using SQOOP and processing data like importing and exporting of databases to the HDFS.
* Experienced the integration of various data sources including Java, JDBC, RDBMS, Shell Scripting, Spreadsheets, and Text files.
* Hands on experience in ingesting data into Data Warehouse using various data loading techniques.
* Developed python scripts to load data to hive from HDFS.
* Participated in developing ETL components for executing various workflows
* Developed pig and hive scripts for processing the data.
* Developed the data lake with marketing analytics data.
* Scheduled Jobs using crontab, rundeck and control-m.
* Importing and exporting data job's, to perform operations like copying data from HDFS and to HDFS using Sqoop.
* Data was ingested which is received from various database providers using Sqoop onto HDFS for analysis and data processing.
* Wrote and Implemented Apache PIG scripts to load data from and to store data into Hive.
* Managed the imported data form different data sources, performed transformation using Hive, Pig and Map- Reduce and loaded data in HDFS.
* Executed Oozie workflow engine to run multiple Hive and Pig jobs, which run independently with time and data availability.
* Created User Defined Functions (UDF’s) for maintaining Incremental ID’s.
* Used Shell scripting to analyze the data from SQL Server source and processed it to store into HDFS.
* Had good exposure to spark Mlib, Streaming and sql by closely working with data scientists.
* Generated reports from Hive data using Micro strategy.

**Environment:** Hadoop, Hive, Sqoop, pig, Python 2.7/3, Java, Flask, MySQL, MS SQL Server, Linux, Shell Scripting, SQL.

FM Global -Johnston, RI Feb 2016 to June 2016

Python Developer

This is a [mutual insurance](https://en.wikipedia.org/wiki/Mutual_insurance) company, with offices worldwide, that specializes in loss prevention services to large corporations throughout the world in the Highly Protected Risk(HPR) property insurance market sector.

Responsibilities:

* Participate in requirement gathering and analysis phase of the project in documenting the business requirements by conducting workshops/meetings with various business users.
* Worked with team of developers on Python applications for RISK management.
* Developed Python application for Google Analytics aggregation and reporting.
* Used flask configuration to manage URLs and application parameters.
* Worked on Python Open stack API's and used Python scripts to update content in the database and manipulate files.
* Configured ec2 instances and configured IAM users and roles.
* Created s3 data pipe using Boto API to load data from internal data sources.
* Developed UDF's to provide custom hive and pig capabilities.
* Performed Data transformations in HIVE and used partitions, buckets for performance improvements.
* I have used Pandas API to put the data as time series and tabular format for east timestamp data manipulation and retrieval.
* Added support for Amazon AWS S3 and RDS to host static/media files and the database into Amazon Cloud.
* Systems automation utilizing Control-M for scheduling and PowerShell/C# for script development.
* Used Pandas library for statistical Analysis.
* Developed tools using Python, Shell scripting, XML to automate some of the menial tasks. Interfacing with supervisors, artists, systems administrators and production to ensure production deadlines are met.
* Also used Bootstrap as a mechanism to manage and organize the html page layout.
* Wrote and executed various MYSQL database queries from Python using Python-MySQL connector and MySQL dB package.

**Environment:** Python, Hive, Oozie, Amazon AWS S3, MySQL, MS SQL Server, Linux, Shell Scripting.

**Deutsche bank – NJ** **Feb 2015 to Nov 2015**

**Python Developer**

Deutsche bank is a German global banking and financial services company offering financial products and services for corporate and institutional clients along with private and business clients. Services include sales, trading, research and origination of debt and equity; mergers and acquisitions (M&A); risk management products, such as derivatives, corporate finance, wealth management, retail banking, fund management, and transaction banking.[[](https://en.wikipedia.org/wiki/Deutsche_Bank#cite_note-8)

**Responsibilities:**

* Worked with team of developers on Python applications for RISK management.
* Designed the database schema for the content management system.
* Designed and developed the UI of the website using **HTML, XHTML, AJAX, CSS** and JavaScript.
* Involved in development of Web Services using **SOAP** for sending and getting data from the external interface in the XML format.
* Wrote Python routines to log into the websites and fetch data for selected options.
* Performed testing using Django's Test Module.
* Skilled experience in installing, configuring, and using Apache Hadoop ecosystems such as Pig and Spark.
* Built the entire Hadoop platform from scratch.
* Experience in ingesting real time/near real time data using Flume, Kafka, Storm
* Evaluated suitability of Hadoop and its ecosystem to the above project and implementing / validating with various proof of concept (POC) applications to eventually adopt them to benefit from the Big Data Hadoop initiative.
* Estimated the Software & Hardware requirements for the Name Node and Data Node in the cluster.
* Extracted the needed data from the server into HDFS and Bulk Loaded the cleaned data into HBase using MapReduce
* Written the Map Reduce programs, Hive UDFs in Java.
* Develop HIVE queries for the analysts.
* Created an e-mail notification service upon completion of job for the particular team which requested for the data.
* Defined job work flows as per their dependencies in Oozie.
* Closely observed building the Reporting Application, which uses the Spark SQL to fetch and generate reports on table data
* Knowledge in performance troubleshooting and tuning Hadoop clusters in Cloudera
* Worked on middle tier and persistence layer. Created service and model layer classes and Value objects/**POJO** to hold values between java classes and database fields.
* Exported/Imported data between different data sources using SQL Server Management Studio.   
  Maintained program libraries, users' manuals and technical documentation.
* Responsible for debugging and troubleshooting the web application.
* Successfully migrated all the data to the database while the site was in production.
* Implemented the validation, error handling, and caching framework with Oracle Coherence cache.
* Worked on scripts for setting up the discovery client with attribute data. Worked on scripts(granite reference data scripts) for setting up adapter attributes in granite system.

**Environment:  Python 2.7, Hadoop, Django 1.4, HTML5, CSS, XML, MySQL, JavaScript, jQuery, Mongo DB, MS SQL Server, JavaScript, GitHub, AWS, Linux, Shell Scripting, AJAX**.

**Efftronics Systems Pvt. Ltd– Vijayawada, India July 2013 to Dec 2013**

**Internship**

Efftronics System pvt ltd (Estb 1985), IT COMPANY nurtured with the latest technological trends, offers unique solutions in all aspects of Information Technology. The services involve Data Collection, Data Validation, Data Processing, Data Communication and Data Dissemination. Comprehensive Design & Development capabilities under popular platforms and the absolute commitment to growing challenges make us the primary choice in the Integration of Software, Graphics and Hardware and embedded systems.

Responsibilities:

* Involved in various phases of Software Development Life Cycle (SDLC) such as requirements gathering, modeling, analysis, design and development.
* Generated Use case diagrams, Activity flow diagrams, Class diagrams and Object diagrams in the design phase.
* Responsible for entire data migration from Sybase ASE server to Oracle
* Migration of API code written for Sybase to Oracle.
* Overlook the migration activity of **PL/SQL** programs
* Migration of the PL/SQL code from Sybase to Oracle.
* Migration of the data contained in the earlier **ASPL** Database from Sybase to Oracle.
* Migrate the Libraries written using Sybase API's to Oracle's OCCI API's
* Automation of testing using Python.

Environment: Python 2.7, Shell scripting, PL/SQL, SVN, Quality Center, Windows.

EDUCATION DETAILS:

* Masters in electrical engineering, Texas A&M University- Kingsville.
* Bachelors in Electronics and Communication Engineering, Vignan University, Guntur, India.