Thulasitharan Govindaraj

Associate - Projects

Big-data engineer with 4+ Years of Experience in the IT Sector and have a strong desire to utilize my skills in a project team and gain more knowledge, which will be useful for the growth of the Organization and also for personal development.

✉ thulasitharan.gt96@gmail.com

☎ 8015811656

stackoverflow.com/users/11430389/

in linkedin.com/in/thulasitharan-govindaraj-2b42651

github.com/ThulasitharanGT

CARRER SUMMARY

- Brought up in Big Data – Hadoop Ecosystems which includes Spark/SCALA , Hive, Impala ,kafka and with diverse proof of concepts hands-on.
- Developed SCALA UDF's and used in it Spark SQL and Spark DataFrame's .
- Performance tuning ,troubleshooting and optimization in Spark.
- Tracing and backtracking the root cause of issues (Finding RCA's).
- Deploying code in PROD Environment and wrote DQC's for the whole project Flow.
- Working experience on creating, Transforming/manipulating DataFrames and DataSets in Spark.
- Vast experience in Spark functionality and programming.
- Worked in data Ingestion and data computations environment.
- Performing updates on PROD environment files which involves critical file changes and validating changes after applying.
- Worked with identifying different schema records in same files identifying and splitting the relevant schema record's, data pipelining and schema evolution handling.
- Worked on developing applications for AWS EMR and Cloudera CDH clusters.
- Worked on Spark streaming as well as batch application.
- Worked on integration between Spark and snowflake, Spark and kafka, Spark and HBASE , spark and KUDU and development of applications for the same.
- Worked on Snowflake data warehousing and have hands on in Snowsql .
- Have worked in Amazon Deequ a DQC(Data Quality Check) Framework on Spark, which can check data for correct value's and validate data quality.
- Have worked on databricks notebook environment and transferred knowledge about the same for whole team.
- Have hands on experience in kubernetes container orchestration engine and Rundeck job scheduler.
- Have written complex flattening and deep flattening functions for dataframes with complex data types such nested array(struct) and struct (struct) to return all sub columns as a separate column.

Systems Engineer

Tata Consultancy Services

09/2016 – 10/2019 *Chennai*

Achievements/Tasks

Retail - (09/2016 – 10/2019) - Hadoop/Spark developer

- Worked in an Agile DevOps project using scrum methodology, which processes Big Data. Started as a tester and now worked in a combination of development and testing tasks.

P.T.O

- Primary goal of the project is to do data crunching with input files using business logic using Scala code which runs in spark computing engine and using Hive in some modules over the process. This computation is carried out in EMR or CDH clusters with PaaS architecture.
- Worked in performance tuning in spark application’s (broadcasting reference file Dataframe’s, catching/ persisting final Dataframe before action, perfecting number of executor’s and cores, optimizing driver and executor memory) and troubleshooting (finding out where job is stuck due to memory or bad logic and correcting code to handle it).
- Trained 6 employees in business knowledge and technical knowledge on technologies in individual's own understanding.
- First to implement sample API in a demo project in my training phase and understanding its working ,in addition to that also taught the concept for 60 of my batch mates.
- Have lead a team of 5 members on absence of team lead for accomplishing a task and got the task done with help of co-ordination and people skills and got appreciation for the same.
- Worked in Agile Environment which uses JIRA to assign,track,log work and monitor tasks.
- Created DQCs for whole project flow and integrated it along with job which will write the result to a S3 path after it completes check for each module.

Associate

Cognizant Technology Solutions

10/2019 – 10/2020 *Chennai*

Achievements/Tasks

Digital Media- (10/2019 – Present) - BigData Developer

- Working in a Data Ingestion project which involves gathering huge amounts of data from various source, cleansing performing transformations and loading data into a destination.
- Fixing code issues for already existing project and enhancing the code and re deploying it.
- Application migration from HIVE, Shell script, python in On-Prem environment to AWS EMR,EC2,Aroura,Athena using spark in DOCKER and other competitive tasks.
- Processing data in spark using Docker using ECS container and loading it in Snowflake as for end user’s.
- Developing job’s using spark and integrating it inside DOCKER container which computes/cleanses and pushes data to snowflake and have worked on configuring SNOWSQL and using put command and other snowflake command’s in it.
- Identifying different schema records in same files (Schema evolution) identifying and splitting the relevant schema record’s.Working in cloud migration tasks.
- Developed a Streaming application using kafka and Spark Streaming to create a topic and feed data to it, hit topic for data and store it in delta lake Stream table.
- Developed a spark batch application to read data from HBASE and upsert in deltalake table on hourly basis.
- Worked on integration between spark and other application’s(Hbase,Kafka,snowflake,teradata) fixing configuration issues and finding out jar version mismatches to solve integration problems
- Written data quality check code for data using Amazon Deequ DQC framework to validate data for quality (min value, max value , range of values for a column, set of allowed values for a column).
- Worked in Agile scrum methodology which uses JIRA to assign,track,log work and monitor tasks.
- Worked on snowflake warehouse(loading data from internal/external stages, creating snowpipe’s, copy into command’s , creating file format’s etc) and snowSQL a snowflake CLI (pushing data from cluster to table or user stage using PUT,getting data to cluster using GET, stages,copy into etc).
- Created a notebook in DataBricks environment to do a truncate and load of data into snowflake reading from s3.

P.T.O

- Data migration Automation from Redshift to hive and impala tables and validation after migration is done, sending alerts on mail with status of job.
- Written functions which will automatically flatten complex nested datatypes such nested array(struct) and struct (struct) to return all sub columns as a separate column. This was used by many teams across client's projects.

Module Lead

Mindtree Limited

10/2020 – Present *Chennai*

Achievements/Tasks

Hospitality - (10/2019 – Present) - Spark/Scala Developer

- Working in spark/scala streaming/batch development . Developing pipeline for Data Ingestion and computation.
- Reading data from Kafka and processing it and storing it in posgres tables or Hadoop hdfs or AWS s3 or doing some processing over it and then posting that message to another kafka queue for some other application.
- Using amazon EMR cluster to process data using spark. AWS S3 is used as primary storage.
- Using Jenkins to build the code after committing to GIT and deploy it in EMR cluster.
- Using EC2 instances to host Kafka and push and pull data from topic's.
- Manipulating each batch of data which arrives real time in streaming and computing required summary and results and updating the same in posgres.

TECHNICAL SKILLS

- ✧ **Cloud Technology -** Amazon s3, AWS EMR , CDH Clusters, AWS Athena, AWS Glue, AWS RDS, AWS ECS, AWS Cloudwatch, AWS EC2
- ✧ **Computing Engines -** ***Spark***,HIVE (MapR,TEZ), impala
- ✧ **Build Tools -** Maven, Gradle, SBT(POC)
- ✧ **Containerization Framework-** Docker , Kubernetes
- ✧ **Warehouse -** HIVE, Snowflake, Deltalake(Databricks)
- ✧ **Programming Languages -** ***SCALA*** ,JAVA(Code analysis),Python(Code analysis), c(basics),c++(basics),c#
- ✧ **No SQL-** Hbase
- ✧ **BIGDATA File Systems -** ***Hadoop DFS*,** Databrics DFS
- ✧ **Messaging System -** Kafka
- ✧ **Notebooks -** Databrics
- ✧ **DQC Framework -** Amazon Deequ
- ✧ **Databases -** Redshift, Posgres, Netezza, Oracle,Teradata
- ✧ **Database UI Tools -** Dbeaver,DBVisualizer
- ✧ **Cluster Management Tools -** Putty, Winscp, Filezilla
- ✧ **OS -** Windows, Unix
- ✧ **IDE -** ***Intelliji***, Eclipse

CERTIFICATIONS

Course Name	University	Obtained (MM/YYYY)	Course ID
Hadoop 101	**IBM** (Cognitive Class)	*10/2019*	**BD0111EN**

P.T.O

BigData 101	**IBM** (Cognitive Class)	*10/2019*	**BD0101EN**
Spark Fundamentals I	**IBM** (Cognitive Class)	*10/2019*	**BD0211EN**
Spark Fundamentals II	**IBM** (Cognitive Class)	*10/2019*	**BD0212EN**
Scala 101	**IBM** (Cognitive Class)	*10/2019*	**SC0101EN**
Accessing Hadoop Data Using Hive	**IBM** (Cognitive Class)	*10/2019*	**BD0141EN**
Moving Data into Hadoop	**IBM** (Cognitive Class)	*10/2019*	**BD0131EN**
Using HBase for Real-time Access to your Big Data	**IBM** (Cognitive Class)	*10/2019*	**BD0143EN**
Spark Overview for Scala Analytics	**IBM** (Cognitive Class)	*10/2019*	**SC0103EN**
Google Analytics for Beginners	Google Analytics Academy	*01/2020*	Google Analytics for Beginners
Hadoop Pogramming - MAPR & YARN	**IBM** (Cognitive Class)	*04/2020*	**BD0115EN**
Docker Essentials: A Developer Introduction	**IBM** (Cognitive Class)	*04/2020*	**CO0101EN**
Simplifying data pipelines with Apache Kafka	**IBM** (Cognitive Class)	*04/2020*	**BD0123EN**
Introduction to Data Science	**IBM** (Cognitive Class)	*04/2020*	**DS0101EN**
Introduction to Cloud	**IBM** (Cognitive Class)	*04/2020*	**CC0101EN**
Data Science with Scala	**IBM** (Cognitive Class)	*04/2020*	**SC0105EN**
Snowflake Decoded	**Udemy** (E-learning)	*07/2020*	**Fundamentals and hands on Training**
Container & Kubernetes Essentials	**IBM** (Cognitive Class)	*07/2020*	**CO0201EN**
Building Cloud Native and Multicloud Applications	**IBM** (Cognitive Class)	*10/2020*	CC0250EN
SQL and Relational Databases	**IBM** (Cognitive Class)	*10/2020*	DB0101EN
Controlling Hadoop Jobs using Oozie	**IBM** (Cognitive Class)	*11/2020*	BD0133EN
DataOps Methodology	**IBM** (Cognitive Class)	*12/2020*	DE0205EN
Spark MLIIB	**IBM** (Cognitive Class)	*12/2020*	BD0221EN
Data Privacy Fundamentals	**IBM** (Cognitive Class)	*12/2020*	DS0301EN
Bitcoin Introduction	**IBM** (Cognitive Class)	*12/2020*	DS0321EN
Hybrid Cloud Conference – Pipelines	**IBM** (Cognitive Class)	*12/2020*	HCC105EN

Note: Link to certificate's and badge's are present in LinkedIn.
Have cleared **Scala, Java** and **SQL** assessments in **Linkedin**.

ACHIEVEMENTS

- Have received many appreciation mails from clients for finding issues and for providing RCA's.
- Have received many appreciations and thumb's ups on direct status call for completing the task on time , efficiently and perfectly.
- Have received appreciations from higher management for good feedback from client's.
- On the spot award (10/2018 – 12/2018)
  - For finding out a critical issue in a reference file before it got into deployment, fixing it and saving the team and business process from loss.
- Star of the Learners Group (09/2016 – 11/2016)
  - For understanding the concept of API in a short period of time and teaching it to other batch mates.

P.T.O

- Secured a **gold medal** for scoring First rank in the first year of my B.Sc (06/2014)
  - For securing 9.2 GPA in main subjects
- On-Campus placement by TCS by the end of 5th semester , based on merit and various phases of IT and coding skill test's.
- Fixed a prod issue in 3500+ files in a single stretch when was only **1 week** old to the project and got appreciation from **client and higher management** .(10/2019)
- Fixed an issue which involved a problem between Hbase and spark integration on a POC with prod environment and got **client appreciation**.(11/2019)
- First to implement spark with kafka and spark with deltalake in whole client side and got recognized by **internal management** and **client**.(12/2019)
- Implemented Kafka and spark streaming pushing data from kafka to deltalake using spark streaming for a streaming project in prod environment for a test topic and got **client appreciation**.(12/2019).
- First to learn about Databricks platform and share my knowledge,educate other team member's on the same.
- First to learn about Amazon Deequ framework and share my knowledge,educate other team member's on the same.

PERSONAL PROJECTS

Vehicle Showroom Management System (10/2018 – 02/2019)

- An big data analysis implementation project in which SCALA architecture is implemented and triggered from PHP in front end.
- Uses Mysql and Files to read/write the data, implemented in Gradle and Maven build tools.
- Implemented my own framework in the Project, replicating some features from MVC.

Learning project on Kafka (*With*** SCALA and SPARK**(Streaming)*****)(04-2019 - 06/2019)***

- Project involving implementation of Kafka to send and receive messages in topic's. Used SCALA to write code for reading and writing messages to KAFKA.
- Implemented a custom partitioner to send certain data into particular partition of topic alone and read the same.
- Used spark streaming to read the data from topic as a DataFrame and save it in a location so it can be used for further analysis.

Pipeline learning project on Deltalake using Spark and Kafka for Stream input (Data Pipeline) (11/2019 - 01/2020)

- Project involving implementation of DeltaLake to store and process data using SPARK, Two input sources one is an batch source with file's in a location and another is a streaming source using Kafka .
- Primary aim (Batch) of the project is to create a delta lake which will be using Bronze, Silver and Gold infrastructure. Data is pushed to Bronze table by a spark job. Then data in Bronze is read and minor computation is made and pushed into Silver and then to Gold.
- Primary aim (Streaming) of the project is to create a delta lake which will be using Bronze, Silver and Gold infrastructure from a **stream**. Data is pushed to Bronze table by a spark job by reading from a kafka topic. Then data in Bronze is read as a stream and minor computation is made and pushed into Silver .A batch job pushes SILVER data to Gold.
- Project is ran through a shell script containing spark submit's. Shell script name is dynamically passes as Command line argument to another shell script which has a scala class to execute system command's

Learning project on resumable job from failure point (*With*** SCALA and SPARK***)(08-2020-10/2020)***

P.T.O

- Project involving implementation of a E2E pipeline involving taking data from a source , preserving it into a temporary table,appending it to an analytics table and taking snapshot of the entire state of the analytics table.
- Implemented this by creating functions for the same and calling relevant functions for each step which needs tho run for the job state
- Used spark to read data from source and put it into tables.Attached architecture Doc.

Resume from
where it failed last

Learning project on Kafka (*With* SCALA and SPARK and deltalake***)(09-2020- 10/2020)***

- Project involving implementation of Kafka to send and receive messages in topic's. Used SCALA to write code for reading and writing messages to Deltalake table.
- Streaming job reads data and then puts it into a streaming table. Batch job does compaction of the data on daily basis and checks if any extra record has arrived for previous day by compacted bronze vs streaming bronze. If yes triggers a mail else just loads the data into the compacted table.
- After this data is loaded into the aggregated gold table's considering compacted bronze table as silver I terms of delta lake architecture.
- A correction job has been done to load the extra data received in case of past days in bronze vs silver. This only takes the extra records and loads into the silver table. Attached architecture document for reference.

Data Collection
Ingestion Pipeline.

Voting System using c# (10/2015 – 01/2016)

- A windows form application with a good exception handling mechanism and complication free architecture.
- MSSQL 2008 as back-end

PERSONAL BLOG

- Have written a blog about HBASE spark integration,on issue which faced and how I resolved it .
  Link: https://medium.com/@thulasitharan.gt96/spark-hbase-integration-d18efc27af63
- Have written a blog on Amazon Deequ DQC frame work how to use it and showed real time examples for the same.
  Link:https://medium.com/@thulasitharan.gt96/usage-of-deequ-suite-by-amazon-for-dqc-data-quality-checks-for-your-data-de1069f60cea
- Have written a blog on automatic data loading into SNOWFLAKE warehouse through use of snowpipe's (Server-less) from AWS S3 storage system.
  Link:https://medium.com/@thulasitharan.gt96/automatic-loading-of-data-to-snowflake-using-snow-pipe-from-an-external-source-when-new-files-are-4d7cb5ae028c
- Have Written a blog on how to run a spark submit inside docker environment. Can be done by copying jar/.py file from cloud(s3) /local. Explained with local path.
  Link: https://link.medium.com/QGvKEUDZR4
- Have written a blog on how to configure hive's metastore for spark.
  Link: https://medium.com/@thulasitharan.gt96/configuring-hive-metastore-for-spark-2c584e65a52d

P.T.O

EDUCATION

B.Sc Computer Science

Loyola College

06/2013 – 05/2016 *Chennai ,GPA: 9.1*

Higher Secondary Education

Vana Vani Matriculation Higher Secondary School

06/2011 – 04/2013 *Chennai,87%*

P.T.O