

**M Bhaskara Gupta**

**Mobile –732 618 7647**

**Email –** [**basu45019@gmail.com**](file:///C%3A%5CUsers%5CBhaskar%5CDesktop%5CResume%5Cbasu45019%40gmail.com)

**PROFESSIONAL SUMMARY**

A highly motivated and results oriented Big Data and ETL Architect with technical and analytical expertise in Data Integration and Data Warehousing implementations.

* Extensive experience in dimensional modeling, DWH/system architecture, ETL/ELT framework, design, Implementation, best practices, and troubleshooting.
* Extensive experience in maintaining data mapping documents, business matrix and other data design artifacts that define technical data specifications and transformation rules.
* Release management, code deployment across various environment and support/maintenance.
* Ownership of overall deliverables, handle work distribution, time management, prioritizing work and assist in all the supporting tasks to ensure completion of project on time.
* Worked in Waterfall and familiar with Agile methodologies
* Excellent verbal and written communication with ability to effectively put forth ideas at functional/technical level.
* Regular Contributor of ideas and process improvement suggestions by always looking for creative and innovative solutions/new features.
* A proven hunger to learn new technologies and translate them into working solution by performing proof of concepts.

**Technical Qualification:**

* 6 + years of strong demonstrated experience on Enterprise Data Lake Design, Data Modeling and Hadoop technologies – HDFS, Hive, Map Reduce, Tez, Spark with Python/Scala, oozie and Sqoop, Cloud Technologies – AWS,Azure and GCP.
* Have Designed and implemented Enterprise Data Lake for NYL on-premises and cloud .
* Have implemented SCD Type1, Type2, Truncate & Load, Append frameworks using Spark (python and Scala) and Glue.
* Have implemented Spark and Glue frame works to transform XML/JSON and Text files and load into S3 and Redshift.
* Developed Spark scripts by using Python/Scala shell commands as per the requirement.
* Developed Python/Scala scripts, UDFs using both Data frames/SQL/Data sets and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into OLTP system.
* Developed Spark code and Spark-SQL/Streaming to process legacy systems real time data and load into RDBMS.
* Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
* Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself.
* Used Spark API over Hartonworks Hadoop YARN to perform analytics on data in Hive.
* Used Maven for continuous build integration and deployment
* Have implemented ETL jobs using glue and load to S3 and Redshift.
* Have created redshift external and internal tables and make available for data analytics team.
* Have implemented frameworks to ingest daily to cloud using python.
* Implemented Partitions, Dynamic Partitions and Buckets.
* Involved in designing of Conceptual, Logical and Physical Data Modeling.
* Involved in Data Profiling, Attribute Naming Standards and Mapping Sheets based on Business requirements.
* 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.
* Expertise in Hartonworks, Cloudera Hadoop Distribution.
* Have Knowledge in importing and exporting data using Sqoop from HDFS to Relational Database Systems and vice-versa.
* 13 + years of strong demonstrated experience with IBM Infosphere Information Server Tools with excellent grip on
* IBM client tools (Datastage&Qualitystage Designer, Administrator, Director, Admin Console, FastTrack, IA, Business Glossary, Metadata Workbench and ISD).
* Have implemented SCD Type1& Type2 jobs on RDBMS and Hadoop(HDFS/Hive) environment.
* Have implemented jobs using Hive Connector, File Connector, Kafka connector, CFF, Sequential file, Dataset, Vertical Peek, Horizontal Peek, Join, Lookup, Funnel, Filter, Sort, CDC, SCD, Remove Duplicates, Copy, Transformations, Surrogate Key, Oracle Enterprise, Oracle Connector, Db2, Shared containers, Teradata Connector, Netezza Connector, Netezza Enterprises, Web Service and most of all sequence stages.
* Have implemented Datastage server routines to update job related audit information tables and control tables.
* ETL of data to/from heterogeneous technology platforms like HDFS Files, XML Files, major RDBMS platforms, flat files (with header & trailer) and SAP
* Data integration with Mainframes
* Dynamic parameterization using parameter sets and Global parameters, partitioning, performance tuning of Datastage jobs.
* Handled Administration tasks like installations (server, client and Fix packs), create/update/delete users, project deployment, ODBC setup, start and stop services and kill the process ids of jobs.
* Strong understanding of Bi/ Data warehousing architecture, concepts, and ETL strategies
* Change data capture, audit tables, reject processing & data reconciliation, etc.
* Data cleansing, build reference data, data normalization & de-normalization
* Loading slowly changing dimensions (type I, II, III), constraint based loading, target load order, etc.
* Dimensional modeling (star/snowflake schema) & DW BUS architecture, different types of dimensions, fact less fact tables, maintaining hierarchical attributes etc.
* Pre-aggregation of fact tables
* Deep understanding and hands on experience in Oracle with experience in
* Writing medium to complex queries, sub queries, joins
* Working with tables, views, synonyms, indexes, constraints, data types
* Analytical functions, hierarchical queries and aggregate queries
* Deep understanding and hands on experience in Teradata with experience in
* Writing medium to complex queries, sub queries, joins
* Working with Teradata utilities (SQL, B-TEQ, Fast Load, MultiLoad, FastExport)
* Working with tables, views, synonyms, indexes, constraints, data types
* Analytical functions, hierarchical queries and aggregate queries
* Strong knowledge of UNIX Shell Scripting (bash) having experience on
* Directory structure, permissions, NFS mounts
* Major commands, pmcmd, variables, exit status, variable evaluation, processing command line arguments and environment variables
* File operation, text editors, control structures, shell script functions, i/o redirection, logging mechanism
* Process control and monitoring, foreground & background processes
* FTP, NDM, DB connectivity & spooling data
* Autosys Scheduling
* Box jobs, command Jobs, file watcher Jobs
* Setting up job dependencies, starting conditions, date conditions, calendar scheduling
* Creating/updating/deleting Autosys box/command/file watcher jobs by creating and deploying jil (job information language)
* Setting up Autosys virtual machines, machine definitions

**Domain Knowledge:**

* Banking Financial Services – DDA, ML and Trust Data processing.
* Insurance – Client, Policy, Telephony, Service and agents Data processing.
* Manufacturing – Orders, Sales, Bill of Materials, Revenue and Pricing.
* Scientific sector – Tsunami and Metrological services.

**Technical Skills:**

**Hadoop Technologies :** Hortonworks HDP(2.6.5), CDH 5.5.1, Hive, Pig, Sqoop, Oozie, Crunch, Flume, Kafka, Spark(PySpark,Scala), HBase, Impala, Cloudera Hue, Hortonworks Data Flow (HDF), Apache Nifi.

**ETL Tools :** IBM InfoSphere DataStage&QualityStage 11.5.0.2/11.3/9.1/8.7/8.5/8.0/7.5, IBM Business Glossary, Metadata Workbench, FastTrack, ISD and Information Analyzer (IA), Informatica Power Center 9.1/9.6.1, Talend Integration Suite – BigData Edition 6.1/ Talend Open Studio 6.1

**Cloud Technologies :** AWS, Azure, GCP, Redshift , Snowflake

**Databases :** Oracle 10g/11g, Teradata 14.10, Netezza

**Languages :** SQL Programming, UNIX Shell Scripting (bash/ksh), Python, Scala, Autosys JIL, C & C++

**Operating Systems :** UNIX, Suse Linux/Sun Solaris/IBM AIX), Win 9X/XP/7/8

**Methodologies :** Waterfall, Agile

**CI/CD :** Jenkins

**Others :** Dimensional Data Modeling, ER Modeling, SQL Developer, TOAD, Teradata SQL Assistance, WinSCP, Notepad++, Textpad, Autosys r11.3 & JIL, Putty, Git, SVN, Business objects(BODS)

**EDUCATION**

**MCA - Master of Computer Applications, (Pondicherry University)** September 2000 – May 2003

Bharathiar College of Engineering and Technologies, Karaikal.

Aggregate: 74%

**PROFFESIONAL EXPERIENCE**

**Project Title** : **Enterprise Data Management – Customer Experience and CRM**

**Duration :**  Apr 2017 – Till Date

**Client**  **:** New York Life Insurance, Lebanon, NJ **Team Size**: 12

**Role** **:** Big Data and ETL Enterprise Architect/ Tech Lead/ Data Engineer

**Environment** **:** Platform: Python, Spark(pyspark, Scala), Hartonworks Hadoop cluster Ambari, Sqoop, Hive, Nifi, Kafka, AZURE cloud, AWS cloud, Redshift, Redshift Spectrum,S3,EMR,Lambda,SNS,SQS ,Snowflake, Glue, Glue catelog IBM Infosphere Datastage&Qualitystage 11.5 Fix Pack 2 on Linux, Oracle11g, TOAD, Jenkins(CI/CD), JIRA, Rally ,Putty, Autosys r11.3, SQL Developer, SVN, Unix Shell Script, SQL scripts, sqlplus, Unix/Linux, Windows XP/7, Data Modelling with Erwin.

**Description:**

The objective of this project is to design Enterprise Data Lake, integrate the Corporate systems data like Client, Policies, Agents, Telephony, Service, Asset, User and PCS data to Data lake (Cloud and Hadoop Clusters) for analytics purpose and push the required data to DAL tables which used by Sales force CRM team for client services and also push the data to IVR search engines, Compliance through Nifi and Kafka.

This project implementing following functionalities:

* Designed and Implemented Enterprise Data Lake.
* Prepared process, standard and Best practices documents for NYL Enterprise Data Lake.
* To process and integrate Client, Agents, Asset, User and PCS data coming from Corporate system into Enterprise EDM Data Lake.
* Implemented ETL complex jobs like SCD Type2, Type1 and Refresh load for client, agents, address, phones, emails, assets, etc.,
* Implemented jobs to push required delta data to IVR search engines, Analytics Vendors and compliance applications using Kafka clusters.
* Creation of DAL table to be used by Salesforce CRM application users for client service.
* Creation of controls to make sure all daily corporate data being loaded into Data Lake.
* Migrated on premises EDM Data Warehouse to Cloud services like S3 ,Redshift, Redshift Spectrum, Glue catalog and Snowflake.

**Job Role:**

* Played a role of Big Data,Cloud and ETL Architect working directly with Line of Business Owner on Data design and sourcing to be able to leverage Data Lake capabilities to scale applications for future advance analytics and make it as centralized data source for all business applications which built on Hortonworks Big Data platform using Spark (pyspark, Scala), Redshift, Redshift spectrum, S3, Postgress and IBM Big Integrate.
* Implemented SCD Type1, Type2, Append, Truncate and Load framework on Python Spark (pyspark, Scala), IBM Big Integrate which integrate the data to Data Lake.
* Developed Spark scripts by using Python/Scala shell commands as per the requirement.
* Developed Python/Scala scripts, UDFs using both Data frames/SQL/Data sets and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into OLTP system.
* Developed Spark code and Spark-SQL/Streaming to process legacy systems real time data and load into RDBMS.
* Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
* Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself.
* Used Spark API over Hartonworks Hadoop YARN to perform analytics on data in Hive.
* Used Maven for continuous build integration and deployment
* Implemented a framework to stream the data from Data lake to downstream systems/Applications like Azure/AWS cloud, Traditional DB’s, IVR, Compliance, Verint and Call Miner using ETL, Nifi, Kafka and SFTP.
* Involved in Data Modeling like understanding source data and preparing STM’s.
* Automating the process of creating and ingesting data on the Data Lake with a single generic framework for all different data source types reducing the maintenance overhead from 10,000+ code streams to 4 code streams.
* Guiding the teams on Automating the process of Creating tables on Hive for 10000+ tables
* Design and implement data governance strategies over Data Lake and suggest best practices for Hadoop Big Integra framework across environments on Dev/QA/prod.
* Develop a reconciliation framework to bring in audit controls for data.

**Project Title** : **Enterprise Data Management (EDM) - US/ML Trust Sourcing**

**Duration :** Mar 2015 – Apr 2017

**Client**  **:** Bank of America, Hartford, CT **Team Size**: 12

**Objective** **:** Data integration with DW for client decision making reports.

**Role** **:**  Project/Tech Lead and Developer

**Environment** **:** Platform: Spark with Python(pyspark), CloudEra Hadoop Cluster, IBM Infosphere Datastage&Qualitystage 11.3 on Linux, Teradata 14.10, TOAD, Putty, Autosys r11.3, SQL Developer, SVN, Unix Shell Script, SQL NDM, BTEQ scripts, sqlplus, Unix/Linux, Windows XP/7

**Description:**

US Trust and Merrill Lynch (ML) are acquired by BOA and currently 5 different SORs (US Trust, DDA, and ML & BOA) are loaded to multiple tables. The USTrust and ML stream should be merged to DDM along with the required changes for SOR fields coming from Authorized data provision point called GDPP. Field level changes like transformations on equivalent fields, removal or addition of necessary fields based on the existing fields in DDM are to be done.

This project implementing following functionalities:

* To process and integrate US Trust and Merill Lynch customer transactions data coming from GDPP in to EDM DDM Hadoop and Teradata environment.
* Implemented ETL complex jobs like SCD Type2 and Refresh load for US and ML Trust account data.
* Creation of ATLAS output files to be used by ATLAS users for analysis and Decision Making.
* Creation of controls to make sure all daily customers data being loaded into DDM.

**Job Role:**

* Big Data Architect, Project/Tech lead and Developer, Leading the team of 8 Data Engineering development team and 4 testing team.
* Gather project requirements from client.
* Interacted with End users to understand the business requirements and prepared Source and Target Mapping(STM’s) documents.
* Experience in leading a team of people (Onsite and offshore) during the Project life cycle.
* Managed Data stage environments. (DEV, UAT & Production).
* Prepared SRS, HLD and LLD documents based on client requirements.
* Prepared project architecture document presented/reviewed with client ETL architecture to get approvals to proceed further implementation.
* Prepared Technical Specification Documents.
* Prepared Unit Test Cases, System Integration Test Cases and Performance Testing documents.
* Resolving the client issues spontaneously without any delays.
* Keep on updating the current issues with client.
* Assign the tasks to the team and maintain the status of assigned tasks using DML Tracker.
* Preparation of Status reports on daily basis.
* Developed complex ETL jobs using Spark with Python(pyspark) and Datastage.
* Developed Spark scripts by using Python/Scala shell commands as per the requirement.
* Developed Python/Scala scripts, UDFs using both Data frames/SQL/Data sets and RDD/MapReduce in Spark for Data Aggregation, queries and writing data back into OLTP system.
* Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* Optimizing of existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frames and Pair RDD's.
* Experienced in handling large datasets using Partitions, Spark in Memory capabilities, Broadcasts in Spark, Effective & efficient Joins, Transformations and other during ingestion process itself.
* Load the data into Spark RDD and performed in-memory data computation to generate the output response.
* Used Spark API over CloudEra Hadoop YARN to perform analytics on data in Hive.
* Used Maven for continuous build integration and deployment
* Have involved build jobs, scheduling jobs through AutoSys and deployment using TAC.
* Extensively used Unix Hadoop commands to files push, get and display, etc.,
* Assign schemas and create Hive tables.
* Loading the processed Hadoop data files into Hive tables.
* Implemented complex mappings such as Slowly Changing Dimensions with Change Data Capture (CDC) stage.
* Data profiling is done with the existing tables before getting it from the source systems.
* Developed the reconciliation jobs to match the record counts in source system with the target data.
* Experienced with Quality stage for data profiling, standardization, matching and survivorship.
* Designed and developed the jobs for extracting, transforming, integrating, and loading data using DataStage Designer.
* Designed the ETL processes using DataStage to load data from Mainframe, XML, Fixed files Flat Files.
* Designed and Implemented audit related flow, tables and integrated in this project.
* Prepared project Deployment and Back out plan documents.
* Responsible to create IDD document for each Datastage job flow and review the same
* Performed the DataStage admin activities for the project.
* Extensively worked on Error Handling.
* Good understanding of Data warehouse concepts and Star Schema and Snowflake Schema model techniques.
* Used the DataStage Director to run, monitor, and test jobs on development and to obtain the performance statistics.
* Developed Unix Shell Scripts to get the source files, to execute ETL jobs, Purging and Archiving the data.
* Involving in taking care of versioning management of code and documents using SVN and BOA SharePoint.
* Involving in creation of Autosys JILs and scheduling the jobs.
* Involving in SIT and UAT support.
* Involved in IQA and EQA review of code and test cases.
* Involved in Performance Tuning of Jobs using Performance Statistics.
* Involved in the production implementation and post implementation support issues.
* Developed BTEQ scripts to load/update the data into target tables.

**Project Title** : **PSC (Parts structure and costing) Duration**: Jan2013–Feb2015

**Client**  **:** JAGUAR LANDROVER LIMITED, UK **Team Size**: 5

**Objective** **:** Integrating the vehicle parts price data with Data warehouse after costing.

**Role** **:** Informatica Developer, DB Development & Shell Scripting

**Environment** **:**  Informatica Powercenter9.1, Oracle 11g, TOAD, SQL Developer, Putty, Unix Shell Script, PL/SQL, sqlplus, Suse Linux, Windows XP/7, BODS.

**Description:**

The objective of this project is to implement the interfaces to perform the costing of produced vehicle and part in the JLR plants based on VBOM, Structure, COA External part prices, Internal transfer prices, etc., and it performs the costing of Daily and EOP transaction costing.

**Project Title** : **Enterprise Data Management Service, POSS R3, O/V Pricing**

**Duration** : June2011–Dec 2012

**Client**  **:**  General Motors Holding LLC, Detroit, MI **Team Size**: 4

**Objective** **:** Installation and testing functionalities of IBM IIS8.5 Tools

**Role** **:**  Project Leader and Developer.

**Environment** **:**  IBM Infosphere Information Server8.5onLinux, Oracle 11g, Putty, SQLDeveloper, SVN, UnixShellScript, sqlplus, Unix/Linux, WindowsXP/7

**Description:**

Enterprise Data Management Service (EDMS) involves the deployment of pre-selected COTS products called IBM InfoSphere Information Server Foundation Tools into the Global Integration Factory (GIF) environment.

**Project Title** : **Indian System Early Warning for Tsunami and Storm Surges**

**Duration** : June2007–May 2011

**Client**  **:** National Centre for Ocean Information Services (INCOIS) **Team Size** : 15

**Objective** **:** Data Migration and Integration

**Role** **:**  ETL Developer.

**Environment** **:** Ascentail Datastage 7.5.xonLinux, Oracle 11g, Putty, SQLDeveloper, SVN, Unix Shell Script, sqlplus, Unix/Linux, WindowsXP/7

**Project Title** : **Traffic Server (web proxy and caching solution)**

**Duration**: Jun2005–Jun2007

**Client**  **:** Inktomi Corp (Yahoo), CA, USA **Team Size** : 10

**Objective** **:** To bring different sets of Information from different lines of business and store in a centralized database for MIS reporting and analysis

**Role** **:**  Developer

**Environment** **:**  C and C++, UNIX (Sun Solaris/IBM AIX), Windows XP