| **Praveen Ganghishetti**  *Bigdata Architect*  H1B Consultant of VCIT Solutions, Inc.  Contract Positions only  [marketing@vcits.com](mailto:marketing@vcits.com)  Tel: (678) 707-2060 |
| --- |

**PROFESSIONAL SUMMARY**

* 16+ years of IT experience in Analysis, Design, Development, and implementation in Big Data and Data warehousing technologies using Big Data Hadoop, Spark, Python, Java, Informatica, Oracle SQL, PL/SQL, and Teradata.
* Good work exposure on Cloudera and Hortonworks distribution platforms.
* Solid hands-on development in Hadoop technologies, including Spark using Python and Java, Hive, Impala, Sqoop, and AWS EMR.
* Extensive experience working with Spark distributed Framework involving Resilient Distributed Data sets (RDD) and Data Frames (Spark SQL) using Python, Java8, Scala.
* Good Experience in using Pandas and Numpy libraries in Python.
* Used Java 8 to Create Spark RDDs and Spark Datasets.
* Used Spark SQL to read and write with external RDBMS like MySQL and Oracle.
* Built a transactional Layer on top of AWS S3 by using Apache Hudi in the AWS Lakehouse.
* Strong knowledge in OLAP Systems, Kimball and Inmon methodology models, Dimensional data modelling using Star schema and Snowflake schema.
* Built Data Governance Framework to define Data Ownership, Data Quality Management, Compliance and Security, and Metadata Management.
* By enabling Data Governance, which is used to define Data Classification, Data Access& Usage, Data privacy Compliance, and Data Retention Purging.
* Used Data Governance tools for Data Catalogs, Data Quality Checks, Access Control, and Regulatory Compliance.
* Experience in Extraction Transformation and Loading (ETL) data from various data sources into Data Marts and Data Warehouse using Informatica PowerCenter components (Repository Manager, Designer, Workflow Manager, and Workflow Monitor).
* Extensively worked on Teradata, Primary, Secondary, Partition Primary, and Join Indexes.
* In-depth expertise in the Teradata cost-based query optimizer, identified potential bottlenecks with queries from the aspects of query writing, skewed redistributions, join strategies, optimizer statistics, physical design considerations (UPI, NUPI, USI, NUSI, PPI, and JI).
* Expertise in building scalable Snowflake data warehouses, implementing complex ELT pipelines, and integrating with modern cloud ecosystems (AWS, Azure, GCP).

**Education**

* Master of Technology in Information Technology, Hyderabad Central University (HCU), India.
* 2011 Bachelor of Technology in Computer Science, JNTU, India. 2007

**Technical skills**

**Languages** Python, Java, SQL, PL/SQL, T-SQL, Unix Shell Scripting, Java, Hadoop, MapReduce, Hive, Sqoop, Spark Core, and Spark SQL using Python, Java8, Scala, Amazon EMR, Spark Streaming, Kibana, SOLR, Apache Elastic

**RDBMS** Oracle 8i/9i/10g, Teradata V2R5/V2R6/12, MS SQL server 2008, DB2, MySQL

**Tools**: Apache Elasticsearch, Apache **SOLR**, Lire **SOLR**, SVN, Rally, DB Visualizer, Teradata SQL Assistant, TOAD, SQL Plus, SQL Developer, MS Visio, JIRA, HP Quality Center, Citrix, Mercury Quality Center, Erwin

**SCM Tools:** Harvest, Kintana, Informatica Version Control, PVCS

**Data Warehousing:** Informatica Power Center 9.6(Designer, Repository Manager, workflow manager, Workflow monitor), Informatica DT Studio

**Distributions:** Hortonworks Data Platform, Cloudera

**Professional Experience**

**Client: Bank of America (Akkodis), Atlanta, Georgia**

**July 2024 – Present**

**Role: Senior Data Architect**

**Responsibilities**

* Work with the Business Analysts to understand the requirements and work on a high-level level detailed design to address the real-time issues in production.
* Develop and implement a comprehensive Data Governance framework to ensure data quality, security, and compliance.
* Work with the Information Architecture team to propose technical solutions to business problems
* Identify gaps in technology and propose viable solutions
* Take accountability for the technical deliveries from offshore
* Understand Hadoop, Spark, Python, other ecosystems like Impala, Hive, Oozie, Pig,, etc., Autosy,, and UNIX Shell Scripting.
* Work with the development teams and QA during the post-code development phase.
* Used Data Governance Framework to ensure data accuracy, Completeness,, and Consistency
* Identify improvement areas within the application and work with the respective teams to implement the same.
* Involved Data Governance Framework in data-related legal and regulatory requirements
* Ensuring adherence to defined process quality standards, best practices, and high-quality levels in all deliverables.
* Adhere to the team’s governing principles and policies.
* Strong working knowledge of ETL, database technologies, big data,, and data processing skills.
* Developed complex Snowflake stored procedures and tasks to automate ELT workflows.
* Integrated Snowflake with AWS S3 using external stages and secure data sharing.
* Built and optimized DBT models to transform raw data into a curated layer for business reporting.
* Led the optimization of compute and storage costs by implementing clustering keys and warehouse monitoring.

**Client: BCBSA (Savi Technologies Inc), Atlanta, Georgia**

**May 2022 –July 2024**

**Role: Big Data Architect**

**Responsibilities**

* Designed scalable and resilient data architectures on AWS to handle large volumes of data efficiently.
* Ensured that data architectures are scalable and highly available to handle growing data volumes and meet SLAs (Service Level Agreements) for availability and performance.
* Planned and implemented data ingestion pipelines to collect data from various sources into AWS services such as Amazon S3, Amazon Kinesis, or AWS Glue.
* Establish Data Governance policies, procedures, and best practices to standardize data management across the organization.
* Create custom analytics and data mining algorithms to help extract knowledge and meaning from vast stores of data.
* Designed and implemented data processing solutions using AWS services such as Amazon EMR (Elastic MapReduce), AWS Glue, or AWS Lambda to transform and analyze data at scale
* Integrating data from different sources and formats, including structured and unstructured data, and ensuring data consistency and integrity.
* Optimizing data processing and analysis workflows for performance and cost-efficiency, including tuning AWS services configurations and selecting appropriate instance types.
* Define and enforce Data Governance standards to ensure consistency, accuracy, and reliability of data assets.
* Documenting data architectures, processes, and best practices and sharing knowledge with team members through documentation, presentations, and training sessions to foster a culture of learning and knowledge sharing within the organization.
* Integrating data from various sources, including databases, data warehouses, streaming sources, and external APIs, to provide a unified view of data for analysis and processing.
* Built a transactional Layer on top of AWS S3 by using Apache Hudi in the AWS Lakehouse.
* Used Redshift Spectrum to enable a unified SQL interface that can accept and process SQL statements where the same query can reference and combine datasets hosted in the data lake as well as data warehouse storage.
* Collaborate with business and IT teams to align Data Governance strategies with business objectives and regulatory requirements.
* In the AWS LakeHouse Architecture, the data warehouse and data lake are natively integrated at the storage as well as common catalog layers to present a unified Lake House interface to processing and consumption layers.
* Optimizing data processing and storage systems for performance, scalability, and cost-effectiveness, including tuning parameters, optimizing queries, and selecting appropriate hardware or cloud resources.

**Technologies:**

Spark 2.3, Python with Spark, with Spark, Amazon S3, AWS Glue, AWS Athena, Boto3, MySQL, Oracle 10g, Informatica Power Center, Elasticsearch, Apache **SOLR**, Lite **SOLR**, DB Visualizer, SharePoint 2010, Rally

**Client: CSRA, Falls Church (Savantis Solutions LLC), Virginia**

**January 2019 – February 2021**

**Role: Lead Big Data Developer/Architect**

**Responsibilities**

* Designed, developed, and implemented Big Data analytics solutions on a Hadoop–based platform.
* Refine a data processing pipeline focused on unstructured and semi-structured data refinement
* Create custom analytics and data mining algorithms to help extract knowledge and meaning from vast stores of data.
* Supported quick turn and rapid implementations and larger scale, and longer duration analytic capability implementations.
* Lead Data Governance initiatives to improve data stewardship, data lineage, and data cataloging.
* Created Spark SQL and Spark RDDs using Java 8 as part of the Hortonworks HDP 3.1 migration.
* Configure the Data flows from different sources (relational databases, XML, JSON) and orchestrate them using Nifi.
* Developed Spark Frameworks using Pyspark and Java to build Raw/Analytical Layers in Big Data.
* Developed utilities in Python and Core Java.
* Write the Data Extraction, Processing,, and Transformation scripts using HIVE and SPARK wherever needed.
* Used Jenkins for Continuous Integration and Git for Version Control.
* Implement Data Governance tools and technologies to automate data quality monitoring, metadata management, and data classification.
* Wrote shell scripts and job management scripts to invoke and manage the Data Ingestion steps.
* Designed HIVE tables for better performance,, as the data volume would be very high. Apply partitions wherever needed.
* Designed and developed the SPARK programs that process high volumes of data with higher processing speeds.
* Work on AWS Services like S3, EMR, Lambda, Glue Jobs, and Athena as part of the Open Data initiative.
* Created RedShift Tables with Various Distribution Styles such as ALL, AUTO, KEY, EVEN.
* Develop and monitor key Data Governance metrics and KPIs to assess data integrity and compliance.
* Created Redshift External Schema to Postgres Database
* Grant Access on Glue Data Catalog to Redshift Cluster.

**Technology**

Java8, Spark 2.3, Python with Spark, with Spark, Hadoop, Hive, HDP 2.6, Amazon EMR, Amazon S3, AWS.

Lamba, AWS Glue, AWS Athena, Boto3, MySQL, Oracle 10g, Elastic Search, Apache SOLR, Lite SOLR, DB Visualizer, SharePoint 2010, Rally, SVN, Maven, GIT, Jenkins.

**Client: Wellmark (Savantis Solutions LLC), DeSimone’s, Iowa**

**November 2018 – January 2019**

**Role: Data Lead**

**Responsibilities**

* Involved in the full life cycle of the project from Design, Analysis, logical and physical architecture modeling, development, Implementation, and testing.
* Responsible for the design and creation of Hive tables, partitioning, bucketing, loading data, and writing Hive queries. Implemented, migrated the existing Hive Script in Spark SQL for better performance.
* Developed Spark streaming jobs in JAVA 8 to receive real-time data from Kafka, process and store the data to HDFS.
* Acted as a single POC between business and technical teams for data and reports.
* Conduct regular Data Governance audits and assessments to identify gaps and implement corrective actions.
* Involved in designing a system that would process 1B records a day and store 7 years of such data, which would involve Hive, Sqoop, Spark, Kafka, Java UI, Presto, and Oracle. This system involved a massive size and number of Tar files and binary files.
* Gather and create business requirement documents and be involved in framing the logic for extraction, transformation, and loading (ETL) processes.
* Created technical design and documented system and technical process changes per Business requirements.
* Promoted code and provided warranty in the higher pre-production and production environments.
* Prepare Korn shell (Unix/Linux) scripts and integrate with the scheduler for automating the Informatica jobs.
* Ensures the integration of software packages, programs, and reusable solutions on multiple platforms.
* Define roles and responsibilities within the Data Governance framework, ensuring clear accountability across teams.
* Coordinate back-out plans for test and production environments.
* Performance tuning of the Teradata queries and help with data modelling in the EDW (Enterprise Data Warehouse) stream.
* Follow all steps of the system life cycle and project SDLC phases in all technical work.
* Determine the best logical and physical architecture for the project and maintain the architectural integrity of the software.
* Identify Root cause analysis and resolve testing defects. Assist with finding new test data and creating extracts for test data as needed.
* Provide training and guidance to stakeholders on Data Governance best practices and data stewardship responsibilities.

**Technology**

Hadoop, Sqoop, Hive, Cloudera Distribution platform, Spark, Python, Informatica PowerCenter 9.6, Informatica Power Exchange, Mainframes, DB2, Oracle 10g, SQL Developer, SharePoint 2010, Autosys, Jira

**Client: Transamerica (Savantis Solutions LLC), Cedar Rapids, Iowa**

**February 2018 – November 2018**

**Role: Big Data Hadoop Technical Lead**

**Responsibilities**

* Used Spark to land data in Hadoop systems using Python and Scala.
* Developed UDFs, UADFs to prepare the data that will be fed to JAVA MapReduce programs.
* Developed Java code that APIs will use and execute HIVE, PIG scripts as a part of the Java Code.
* Worked on the Continuous Integration of the Big Data build process.
* Moved data to Hadoop to streamline Transamerica’s business by automating many manual processes.
* Pulling data from mainframe policy administration systems and landing the data in Hadoop
* Worked on importing and exporting data, into & out of HDFS and Hive using Sqoop
* Define and implement a scalable Data Governance framework to ensure data accuracy, consistency, and accessibility across the organization.
* Worked on creating Hive tables and wrote Hive queries for data analysis to meet business requirements.
* Extensively used Spark stack to develop a preprocessing job which includes RDD, Datasets, and Data frames APIs to transform the data for upstream consumption.
* Establish data classification and categorization rules to enhance data security and compliance within the Data Governance framework.
* Analyzing the requirement specifications provided by the client and translating them into Technical Impacts on the system.
* Created High-Level design, detailed-level design, design specifications, test plan, and test scripts.
* Involved in code development and code review during development and Integration testing.

**Technology**

Hadoop, Sqoop, Hive, Cloudera Distribution platform, Spark, Python, Informatica PowerCenter 9.6, Informatica Power Exchange, Mainframes, DB2, Oracle 10g, SQL Developer, SharePoint 2010, Autosys, Jira

**Client: American Family Insurance (Infosys), Madison, Wisconsin**

**July 2014 – February 2018**

**Role: Big Data/Informatica Technical Lead**

**Responsibilities**

* Develop and enforce Data Governance workflows for data acquisition, storage, usage, and retirement to maintain data integrity.
* Utilized in-depth knowledge of functional and technical experience in Data Warehousing and other leading-edge products and technology in conjunction with industry and business skills to deliver solutions to customers.
* Wrote Sqoop scripts to import data into Hive/HDFS from RDBMS.
* Developed Hive Queries to analyze the data in HDFS to identify issues and behavioral patterns.
* Used Spark, Python Pandas, and Numpy modules for Data analysis, Data scraping, and parsing
* Implemented Concurrent execution of workflows and Session Partitioning techniques as part of Performance Tuning.
* Worked on Informatica DT Studio to parse input XML and JSON files
* Created Informatica Web service provider using XML Generator, XML Parser, and SQL Transform
* Handled Pushdown Optimization technique to tune the mappings and sessions when working on bulk loads or huge volumes of data.
* Used Debugger in the mappings, identified bugs in existing mappings by analyzing the data flow and evaluating transformations.
* Design and oversee Data Governance policies for master data management (MDM) to ensure a single source of truth for critical data assets.
* Cleaned up mappings to not create lengthy log files by turning off verbose logging and getting rid of warning messages.
* Designed and developed stored procedures using PL/SQL and tuned SQL queries for better performance.
* Implemented slowly changing dimensions (SCD Type 1 & 2) in various mappings.
* Created and used reusable Transformations using Informatica PowerCenter.
* Worked on AutoSys to automate the Execution of Informatica jobs
* Responsible for writing/documenting the Unit Test Cases with different testing scenarios to meet business rules implemented in ETL mappings.

**Technology**

Hadoop, HDFS, Hortonworks distribution platform, PIG, Hive, Python, Spark, Spark SQL, Informatica PowerCenter 9.1, Informatica DT Studio, Oracle 10g, Greenplum, SQL Developer, Winsql, SharePoint 2010, Autosys, Jira, Harvest, XML Spy Editor

**Client: Manulife Insurance (Infosys), Boston, Massachusetts December 2013 - July 2014**

**Role: Sr. Informatica Developer**

**Responsibilities**

* Worked closely with project managers, business analysts, and DBAs to achieve business and functional requirements.
* Worked in Software Development Life Cycle (SDLC), like in Agile Scrum methodologies.
* Used Informatica Power Center 9.1 for extraction, loading, and transformation (ETL) of data in the data warehouse.
* Built efficient ETL Informatica packages for processing fact and dimension tables with complex transformations using type 1 and type 2 changes.
* Lead data governance risk assessments to identify potential vulnerabilities in data handling, storage, and processing.
* Designed and developed complex mappings from varied transformation logic like Unconnected and Connected lookups, Router, Filter, Expression, Aggregator, Joiner, Update Strategy, and more.
* Worked on Informatica Power Center tool - Source Analyzer, Target Designer, Mapping Designer, and Transformation Developer.
* Assisted in the design and Maintenance of the Metadata environment.
* Created Workflows and Sessions to load data from the SQL Server, Oracle, flat file, and XML file sources that exist on servers located at various locations.
* Implement data lineage tracking to provide full visibility into data movement, transformations, and ownership.
* Responsible for creating business solutions for Incremental and full loads.
* Created different parameter files and changed Session parameters, mapping parameters, and variables at run time.
* Created high-level design documents for extracting data from complex relational database tables, data conversions, transformation, and loading into specific formats
* Architect and designed the ETL solution that included designing Mappings and workflows, deciding load strategies, implementing appropriate error handling and error notification processes, scheduling and designing re-usable ETL pieces through parameterization.
* Developed Mapping using parameters, Session parameters, Mapping variable/parameters, and created Parameter files and runs of workflows based on changing variable values.
* Created Unix Shell Scripts to automate pre-session and post-session processes.
* Involved in creating new table structures and modifying existing tables, discussing with the Data Modeler.
* Updated the tables as per the requirements by generating queries using SQL to check the data consistency in the table.
* Created shortcuts for reusable transformation, source/target definitions, and Mapplets in the Shared folder.
* Performed unit testing, Integration testing, Performance, and Functional testing of the mappings.
* Involved in high-level and low-level design, analyzing Source to Target Mapping (STM's), Test Cases, and Code migration reports.

**Technologies**

Informatica PowerCenter 9.6, MS SQL Server 2008, SQL Developer, Toad, T-SQL, Windows Server

**Client: Volkswagen Group of America (Infosys), Detroit, Michigan July 2013 - December 2013**

**Role: Sr. Informatica Developer**

**Responsibilities**

* Identified all OMD/CRM Tables in PRD1, Shared with Other Systems
* Identified all CRM/OMD Tables that Provide Data to Other Systems
* Identified all Interfaces with CRM/OMD Data
* Identified all Tables Referenced by Interfaces
* Catalog all Dependencies on PRD1 CRM/OMD Data
* Formulate Oracle SQL queries on Informatica Metadata database repository to identify Informatica jobs that source/from OMD/CRM systems to other systems.
* Involved in loading analysis tables to SDR-SAMBA’s data repository.
* Involved in the design of SDR-SAMBA’s data repository.
* Analyzed Views, Materialized views, and Stored procedures.
* Analyzed DB audit and Trace to check the Table access, patterns, and frequencies.
* Involved in basic profiling on the DB schema to identify the table data usage and volume.
* Analyzed Informatica Jobs using Informatica metadata to identify the tables involved.
* Analyzed Perl and Unix scripts to identify the tables used.
* Created the Interface dependencies and hierarchies.

**Technologies**

Informatica PowerCenter, Toad, SQL developer, Oracle Metadata, Erwin, Visio, Windows Server

**Client: Cisco Systems (Tech Mahindra), San Jose, California May 2011 - July 2013**

**Role: Informatica Developer**

**Responsibilities**

* Responsible for designing and developing testing processes necessary to extract data from operational databases, transform, and load into the data warehouse using Informatica Power Center.
* Responsible for Modeling, Design, Development, and Integration Testing for the BIDS Platform.
* Make sure that all the Dev and Stage Environments are sanitized within the given SLA.
* Develop the ETLs based on the BRDs and ETL specification documents provided by the client using Informatica.
* Created complex mappings in Power Center Designer using Expression, Filter, Sequence Generator, Update Strategy, Joiner, and Stored procedure transformations.
* Worked extensively on Informatica tools like Designer, Workflow Monitor, and Workflow Manager.
* Worked on all the Transformations like Lookup, Aggregator, Expression, Router, Filter, Update Strategy, Stored procedure, and Sequence Generator.
* Created connected and unconnected Lookup transformations to look up the data from the source to the ETL target tables.
* Wrote SQL, PL/SQL, and stored procedures for implementing business rules and transformations.
* Used the update strategy to effectively migrate data from the source to the target.
* Created test cases and completed unit, integration, and system tests for the the Data Warehouse.
* Involved in debugging and validating the mappings and Code Review to rectify the issue.
* Developed, implemented, and enforced ETL best practices standards.
* Worked on all the Transformations like Lookup, Aggregator, Expression, Router, Filter, Update Strategy, Stored procedure, and Sequence Generator.
* Created Scheduled Sessions and Batch processes based on demand, run on time, and run only once using Informatica Server Manager.
* Re-designed multiple existing Power Center mappings to implement the change request (CR) representing the updated business logic.
* Developed CDC and SCD Type 1,2,3 mappings to meet the business requirements.
* Created reusable transformations to increase the reusability during the development life cycle.
* Responsible for the creation of ETL technical Specification documents based on the BRD.

**Technology**

Informatica PowerCenter 9.1, Teradata V2R5, Oracle 10g, TOAD, SQL Developer, HP Quality Center, Dollar Universe, Kintana, PVCS

**Client: Abercrombie & Fitch (Mahindra Satyam), Columbus, Ohio June 2007 - March 2010**

**Role: Oracle SQL, PLSQL, Forms 10g Developer**

**Responsibilities**

* Involved in developing technical documents based on functional specs.
* Performed data quality analysis to validate the input data based on the cleansing rules.
* Extensively used PLSQL Collections like NESTED Tables and VARRAYS.
* Ensure Quality and On-time delivery.
* Actively participated in gathering Business Requirements and System Specification from system users.
* Analyze the current data management procedures in practice and suggest ways for automating the process or improving the existing system.
* Improved performance and tuning of SQL queries and fixed the slow-running queries in production by using utilities.

**Technologies**

Oracle 11g/10g, SQL, PL/SQL, SQL\*Loader, Oracle Designer, Oracle Forms 9i, Mercury Quality Center.