|  |
| --- |
| **Yagnaraman Madhavan** |

# Experience Summary

* **14 years of professional experience in IT with expertise in Analysis, Design, Development, Implementation, Modeling and Testing**
* **Good balance of technical and management skills, with proven ability to lead data engineering & data science teams and deliver projects on time and on budget**
* **Strong knowledge and experience of AWS cloud services**
* **Strong knowledge and experience in batch and real time data analytics**
* **Strong knowledge and expertise in big data queries and interpretation, exploratory data analysis, predictive modeling, data mining, machine learning algorithms, and KPI dashboards**
* **Good knowledge and experience in developing machine learning model prototypes and deploying the models to production, monitoring model results and model maintenance**
* **Solid understanding and hands-on experience handling various supervised and unsupervised learning algorithms such as linear Regression, Decision trees, Random Forest, Naïve Bayes, SVM and clustering**
* **Worked for 5+ years in various Onsite Assignments in the United States**
* **Hands on Experience in Git and Jenkins, Docker, Kubernetes**
* **Worked on Agile Projects for the past 3 years in scrum and Kanban models**
* Having experience in RDBMS, Data Warehousing, Data lake & Data platform implementation
* Strong programming, analytical, problem solving, communication & Interpersonal skills
* Quick learner and adaptive to new & challenging technological environments.
* Provided design recommendations and thought leadership to sponsors/stakeholders that improved review processes and resolved technical problems
* Collaborates well across technology groups.

# Technical Skills

|  |  |
| --- | --- |
| **Hadoop Distribution** | **Hortonworks, Cloudera** |
| **Analytical Tools/Languages** | **Jupyter, R-Studio, R, Python, Spark ML, Dataiku** |
| **Programing Language** | **Scala, Python, Java** |
| **Cloud Technologies** | **AWS-S3, Redshift, Kinesis, Sagemaker , EMR, Lambda, Cloudwatch Events** |
| **Containerization Technologies** | **Docker, Kubernetes** |
| **Data Scheduling** | **Zena, Control-M, Autosys** |
| **Data Extraction** | **Spark, Kafka, Pig, Hive, Sqoop** |
| **Data Visualization/ BI Reporting** | **Tableau, Hyperion IR, Crystal Reports** |
| **Version Control** | **Git, Gitlab** |
| **Dev Ops** | **Jenkins** |
| **Agile Tool** | **JIRA** |
| **IDE** | **Eclipse, PyCharm** |
| **Data Base** | **Oracle 12C, SQL-Server 2005,Teradata** |
| **ETL Tools** | **Talend, Informatica, Datastage, Powerexchange** |
| **Emulators** | **Putty, Winscp** |
| **Scripting** | **Shell Scripting** |

# Educational Qualification

* **Post Graduate Program in Big Data & Machine Learning** : (PGP-BDML) Oct 2017 – Sep 2018

**Offered by:** Illinois Institute of Technology (Chicago, US), Great Lakes Institute of Management (Chennai, India

* **Bachelor of Engineering: Anna University-**  2002 – 2006

# External Certification

* **Certification Name:** AWS Certified Solution Architect – Associate

**Valid Until: Jul 2022**

**Relevant Project Experience**

|  |  |
| --- | --- |
| **Project Title** | **Social Media Analytics** |
| **Duration** | **Jan 2019 – Current** |
| **Domain** | **Media & Entertainment** |
| **Environment** | **Python, PySpark, AWS S3, AWS Redshift, AWS Lambda, AWS EMR, Hive, Cloudwatch Events, Kibana, Elasticsearch, AWS Sagemaker, Machine learning, Docker, AWS Kinesis, AWS Firehose** |

**Project Description:**

Client is a leading media and entertainment company that is primarily known for hosting wrestling. After each live shows, creative analytics team does social response of the shows: keywords search tagging process for pre-defined properties/topic models tagging process for wrestling programs, storyline models, sentiment models for fan feedback, social conversation volume calculation, clustering & summarization of feedbacks on multiple TV programs to identify key performance indicators which drive business decisions. These decisions help to improve content distribution, understand market sentiments, fan reactions/feedback.

The objective of this program is to ingest tweets data from Gnip vendor into S3 raw zone, create unified data landing tables in Hive, process data from UDL tables, clean, tag, enrich data in EMR cluster using pyspark, run through supervised & unsupervised ML models using AWS sagemaker and load into redshift datawarehouse dimension and fact tables. Data from datawarehouse is then aggregated and loaded into datamart tables for analytical reporting based on requirement metric group. Alerting & notifications is implemented to generate alerts for any abnormal conversations.

**Role: Manager – Data Engineering, Analytics**

**Technical Responsibilities:**

* Build cross functional relationships with data scientists, product managers, and data engineers to understand data needs and deliver on those needs.
* Drive the design, development and implementation of various data pipeline & analytics pipeline functionalities to build an efficient and scalable data platform
* Assist in the definition of software architecture to ensure that the client’s software solutions are built within a consistent framework
* Participate in setting strategy and standards through data architecture and implementation leveraging big data and analytics tools and technologies
* Provide subject matter expertise in the analysis, preparation of specifications and plans for the development of data processes
* Manage development of data resources and support new features
* Manage and develop a team of data engineers, data scientists, BI engineers, providing mentorship and guidance and improve efficiency in process
* Provide scoping, estimation and project planning to a project
* Worked on sprint releases, planning & backlog grooming
* Co-ordinate and manage the monthly revenue forecasting activities
* Manage all custom and adhoc reporting requirements
* Take part in recruiting, including reviewing, shaping, and evolving our process, training team members to interview effectively, and proactively engaging candidates

|  |  |
| --- | --- |
| **Project Title** | **Datalab** |
| **Duration** | **May 2019 – Dec 2019** |
| **Domain** | **Technology** |
| **Environment** | **Python, PySpark, Spark, Scala, Java, Cassandra, Kibana, Elasticsearch, Docker, Kubernetes, AWS S3** |

**Project Description:**

Client’s objective is to build a fully functional data science platform which will enable easier understanding of descriptive/inferential statistics of a dataset, create predictive models and predict using various pre-built algorithms. This will be the platform which will enable various data scientists to easily leverage for executing data science experiments.

Functionalities that will be part of Datalab are:

* 1. Data Discovery & catalog – Centralized Metadata service, enables data discovery
  2. Data Wrangler – Data exploration & wrangling
  3. Notebooks – build rapid prototypes and experiments
  4. Jobs – To schedule or run any tasks
  5. Flows – For wrangling larger datasets
  6. Model Building – Build machine learning models
  7. Model Deployment – Deploy model to test and production environments
  8. Model results Monitoring Dashboards

**Role: Manager - Product Engineering**

**Roles & Responsibilities:**

* Design, develop and improve various product functionalities in scope of application development.
* Lead development of various data science functionalities to build an efficient and scalable platform
* Provide direction and guidance to other product engineers, data engineers to solve problems, improve efficiency and process, and employ new technology
* Provide scoping, estimation and project planning to a project
* Assist in the definition of software architecture to ensure that the client’s software solutions are built within a consistent framework
* Provide subject matter expertise in the analysis, preparation of specifications and plans for the development of data processes
* Participate in setting strategy and standards through data architecture and implementation leveraging big data and analytics tools and technologies
* Co-ordinate and manage the monthly revenue forecasting activities
* Manage all custom and adhoc reporting requirements
* Take part in recruiting, including reviewing, shaping, and evolving our process, training team members to interview effectively, and proactively engaging candidates

|  |  |
| --- | --- |
| **Project Title** | **Data lake - Data Delivery Accelerator Gold** |
| **Duration** | **Dec 2017 - Apr 2019** |
| **Client** | **HCSC** |
| **Environment** | **Hortonworks, Spark, Scala, Kafka, Hive, Hbase, Sqoop, Shell Scripting, Tableau** |

**Project Description:**

HCSC is one of the largest member owned health insurance Company in the United States. HCSC offers group life, disability, and dental solutions, as well as a range of other individual solutions. Data delivery accelerator is a program aimed at uplifting HCSC’s reporting, advanced analytics, and artificial intelligence capabilities and transform into a data-driven insights culture that drives competitive advantage. The program includes: 1. Creation of certified enterprise information assets. 2. Enterprise consumption views. 3. Standard tooling to enable business teams to create reports and advanced analytics.

Data delivery accelerator program includes:

1. Ingestion of batch & streaming data sources from multiple source systems into the data lake raw layer.
2. Integrate data from multiple source systems in raw layer; transform the data sources according to the business transformation rules to create certified data assets in the gold layer.
3. Build consumption views and generate data extracts in the consumption layer after applying the business rules.
4. Augment data by loading data from Data Lake into EDW for reporting.
5. Provision data from Data Lake for enterprise data analytics.
6. Provision data from Kafka messaging layer and NoSQL databases for downstream API’s.
7. Build dashboards for business users for business intelligence reporting.

**Role: Technical Lead – Data Analytics**

**Technical Responsibilities:**

* Involved in the architecture review sessions & made improvements along with the onsite architects.
* Involved in complete Big Data flow of the application starting from data ingestion upstream to Consumption.
* Processing the data in HDFS, analyzing the data, and involved in writing transformation to process data.
* Applied Transformation logic using Spark streaming, Spark Core and Spark SQL to store data in Kafka and Hive.
* Developed Spark code using Scala and Spark-SQL for faster processing of data.
* Handling Hive queries using Spark SQL that integrate with Spark environment implemented in Scala.
* Worked on Hadoop file formats such as csv, Text, Parquet in hive.
* Developed a process for the Batch & real time data ingestion from different sources, after successful completion the files are stored in HDFS & Kafka respectively using Shell Scripting.
* Review of the Scripts and suggesting modification in the scripts.
* Preparing the Unit test plan and executing the Unit testing.
* Built dashboards using Tableau for business intelligence reporting.
* Involved in the POC of a data science project using Dataiku.
* Lead the offshore team, allocated task to team members, and regular update to onsite leads.
* Regular tracking of activities of the team members to meet the time lines.

**Non-Technical Responsibilities:**

* Analysis and Requirement Gathering from the client.
* Coordinating and helping the team technically to meet the deadlines.
* Documented all business processes, needs, flows and diagrams
* Worked with the scrum master in stories planning for the sprints & estimated the story in story points

|  |  |
| --- | --- |
| **Project Title** | **HEDIS** |
| **Duration** | **Oct 2016 - Nov 2017** |
| **Client** | **Aetna** |
| **Environment** | **Hortonworks, Spark, Scala, Kafka, Hive, Sqoop, Shell Scripting, Tableau** |

**Project Description:**

The Healthcare Effectiveness Data and Information Set (HEDIS) is a set of standardized performance measures designed to ensure that purchasers and consumers have the information they need to reliably compare the performance of healthcare organizations. The performance measures in HEDIS are related to many significant public health issues. HEDIS also includes a standardized survey of consumers' experiences that evaluates plan performance in areas such as customer service, access to care, and claims processing. HEDIS is sponsored, supported, and maintained by the National Committee for Quality Assurance (NCQA). HEDIS Team prepares HEDIS reports for the commercial and Medicare HMO and PPO products.

Additionally, HEDIS data is sent to the Network Information for Sales Support (NISS) department. They provide responses to Client/broker/consultant/regulators Employers, brokers, consultants and regulators who are asking health plans to explain their HEDIS results and present improvement plans, whether the results are low compared to Quality Compass norms, have decreased from the previous year’s results or have not been reported.

**Following are the high-level functionalities delivered by this application.**

1. Data Loading – This includes receiving feeds from back offices and loads data from DB2

(EDW) to HDFS on daily/weekly/monthly and AD-HOC basis.

1. Report Generation – This includes scheduling, generating Pre-Canned, and adhoc reports of MA, QA, QISKV, SAMPLING and SURVEY.
2. Report Delivery – This includes delivering the reports to the customers through various channels like Email and sending it to DB2 (EDW) for further comparisons.

**Role: Technical Lead – Data Engineering**

**Technical Responsibilities**

* Involved in Requirement Analysis, Design, Development and Testing of the risk workflow system.
* Responsible for building scalable distributed data solutions using Hadoop.
* Used Spark over YARN to perform analytics on data in Hive.
* Developed Scala scripts using Data frames/SQL/Data sets and RDD in Spark 1.6 for Data Aggregation, queries and writing data back into OLTP system through Sqoop.
* 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 Core, Spark-SQL and Data Frames.
* Experienced in handling large datasets using Partitions and Spark in Memory capabilities.
* Experience in designing and developing POC’s in Spark using Scala to compare the performance of Spark and Hive.
* Involved in setting up the data lake in Hive where the data from DB2 were extracted for report processing.
* Extensively worked in apache Hive for computing Measure Aggregate and Qualifying Info, the two ultimate reports generated from the HEDIS process.
* Involved in importing and exporting of data from DB2 and HDFS using apache Sqoop.
* Involved in writing PIG scripts used to filter the data from the streaming output files, which consist of records with variable length based on their output structure.
* Involved in writing SHELL scripts from which all the PIG, Hive and Sqoop scripts were executed.
* As part of Production support, involved in fixing the failed jobs from Production and Development cycle, root cause analysis and permanent fixes.
* Raised and handled JIRA tickets for the most frequent technical and environmental issues in the jobs.

**Non-Technical Responsibilities:**

* Analysis and Requirement Gathering from the client.
* Coordinating and helping the team technically to meet the deadlines.
* Documented all business processes, needs, flows and diagrams
* Worked with the scrum master in stories planning for the sprints & estimated the story in story points

|  |  |
| --- | --- |
| **Project Title** | **CareXcell Version 1.0** |
| **Duration** | **Sep 2015 - Oct 2016** |
| **Client** | **Cerner Healthcare** |
| **Environment** | **Hortonworks, Spark, Scala, Kafka, Hive, Hbase, Sqoop, Shell Scripting** |

**Project Description**

CareXcell consolidates and normalizes patient data from disparate EHRs and clinical systems into a single patient-centered repository through any health information exchange. Accessing summarized, real-time clinical data improves provider efficiency in identifying and addressing patient-specific issues and risk factors before they become more severe, costly conditions. For example, CareXcell can help providers initiate appropriate interventions as part of patients’ personalized care plans to help improve long-term outcomes, keeping rising-risk patients from becoming high-risk, high-cost utilizers and ultimately strengthening their bottom line*.*

**Role: Lead Developer/Technical Lead**

**Technical Responsibilities**

* Involved in setting up the Patient centric repository database in Hbase such as table creation and loading data into the Hbase master tables using shell script. Manipulation of Hbase tables using various filter queries for data analysis.
* Involved in the enhancements of the map reduce program to export the data from Hbase to SQL server using apache sqoop.
* Involved in writing Sqoop commands for exporting data from Hbase tables to Hive and SQL server for initial as well as incremental load for user report generation.
* Used Hive to analyze the partitioned and bucketed data and compute various metrics for reporting the statistics to the business.
* Involved in writing Oozie workflow XML's that includes java jobs, sqoop jobs, stored procedure and Hadoop commands for the report jobs and other batch jobs. Also written various workflows for scheduling daily, weekly and monthly jobs.
* As part of Production support team member involved in fixing the EVTS tickets, finding the root cause analysis for the major issues reported, scheduling and monitoring jobs.

|  |  |
| --- | --- |
| **Project Title** | **HPXR & EDW** |
| **Duration** | **Aug 2013 – Aug 2015** |
| **Client** | **Excellus Health Plan** |
| **Environment** | **Informatica, Oracle, Teradata, Cognos** |

**Project Description**

The Enterprise Data Warehouse (EDW) is an integrated repository of the enterprise's critical healthcare data and represents this data in a consistent, business-focused manner, in order to meet reporting and analysis requirements across the enterprise.

HPXR is a dimensional data model that stores health plan data and in fact-based subject areas with the intent of reducing data redundancy and providing a simplified data structure to provide a basis for operational reporting to occur and it feeds the enterprise data warehouse with data from a OLTP system called Facets. The Enterprise Data Warehouse (EDW) is an integrated repository of the enterprise's critical healthcare data. The EDW houses primarily Membership and Claims data from our various systems, and represents this data in a consistent, business-focused manner, in order to meet reporting and analysis requirements across the enterprise. EDW will provide a consolidated picture of enterprise data for multiple subject areas, such as Claims, Membership, Products, Providers and Reference Data. It will be a virtual "one-stop-shopping" data environment.

**Role: Senior Developer**

**Technical Responsibilities**

* Performed data profiling and analysis, determined the solution design, reviewed and validated the data model changes.
* Created technical specification documents and conducted reviews within the team.
* Participated in specification reviews and provided inputs to solution design.
* Developed complex Informatica mappings and tuned the performance based on bottlenecks identified from the session log.
* Developed CDC mappings in Power center using Power exchange source as the source definition.
* Managed day-to-day activities of development and ensured that deliverables are completed as per plan.
* Co-ordinated builds and code promotion process.
* Co-ordinated Control-M requests and ad-hoc batch schedule changes.
* Created deployment checklist of informatica objects for post deployment verification and validation.
* Prepared change control request form for production deployment.
* Reviewed code and unit test results.
* Facilitated the ETL inventory file process for maintaining code deliverables.
* Prepare weekly status reports and send it to the client management.

|  |  |
| --- | --- |
| **Project Title** | **Transformation Program** |
| **Duration** | **Jan 2009 – Jul 2013** |
| **Client** | **Excellus Health Plan** |
| **Environment** | **Informatica, Oracle, Teradata, Cognos** |

**Project Description**

Existing Health plan was on mainframe systems initially but a transformation was required in the system with vision of improvement in:

* Customer experience-by simplified process and enhancements
* Simplification in product, platform (use of single platform i.e. Facets) and process.
* Reduction of cost
* To have a centralized system like Facets.
* Lack of IT infrastructure support.
* Reduce operational cost.

The transformation activities include analysis of gaps in the existing system and documentation of the approach to close the gaps and then the actual built of the interfaces, conversions, extensions, bolt-ons, reports, letters and sub-systems, testing & quality assurance. The interfaces are broadly categorized into two types namely Batch Interface and Real-Time Interface. Informatica interface development comes under the scope of batch interfaces. This interface deals with the extraction of the required data from the Facets (Core System)/ HPXR (Replica of FACETS) then the business logics are applied and data cleansing is performed through Informatica and the intended data are finally sent to the Mainframe system through the FTP process. The source and the target for these interfaces could be in Table/Flat Files/XML format.

**Role: Onsite coordinator**

**Technical Responsibilities**

* Analyze business/functional requirements, review the solution design, and develop the technical detailed design.
* Analyze data issues, identify and articulate the business impact of data problems.
* Develop Informatica mappings and test the integration solution individually or in collaboration with onsite/offshore development resources.
* Guide development resources, perform code & test results reviews and ensure that all solutions are aligned to pre-defined architectural standards, guidelines, best practices, and meet quality standards.
* Deploy the solution on a high-availability shared service production environment.
* Understand and comply with the established software development life cycle methodology.
* Establish and enhance technical guidelines and best practices for the development team.

|  |  |
| --- | --- |
| **Project Title** | **Facets Enhancements Operations** |
| **Duration** | **Jul 2008 – Dec 2008** |
| **Client** | **Cigna Inc** |
| **Environment** | **Hyperion IR, Oracle** |

**Project Description:**

Facets Enhancements Operations project involves development/enhancement of service center operation reports, develop Facets Extension application to capture and store Employee details like Employee Type, Business Unit etc... The Extension is built on the Facets EXT database. Data is migrated from Facets EXT database to Facets Daily Deltas & Facets Masters data stores using Oracle CDC. Data is migrated & cleansed from Facets Daily Deltas data store to OneSource History tables using Informatica 8.x. The source data for the report is from OneSource History tables. As part of this project, claims and customer service operational reports are developed. The scope of this project is to develop claims and customer service operational reports and publish the reports.

**Role: Reports Developer**

**Technical Responsibilities**

* Analyzed the report specification document and design of on demand as well as scheduled claims as well as customer service reports using Hyperion Interactive Reporting
* Developed reports based on the user requirements and report design specification document.
* Tested and debugged the reports.
* Published the reports using Hyperion Performance Suite.

|  |  |
| --- | --- |
| **Project Title** | **Data stage Server to Parallel Conversion** |
| **Duration** | **Jan 2008 – Jul 2008** |
| **Client** | **Ingenix** |
| **Environment** | **Data stage Parallel Edition, Oracle** |

**Project Description**

The scope of this project is convert the server version of the existing data stage ETL jobs to its equivalent parallel version. This project-involved conversion of the existing server version of data stage ETL jobs to its equivalent parallel version, unit and performance test to ensure that the parallel capabilities like partitioning is improving the performance.

**Role: Developer**

**Technical Responsibilities**

* Analyzed the business requirements and design specification documents from the design team.
* Designed and developed appropriate parallel ETL jobs to extract and transform data from various sources to meet requirements.
* Performed unit & performance testing of the jobs.
* Monitored jobs using data stage director, analyzed the logs, and made improvements to ensure that the performance is optimal.

|  |  |
| --- | --- |
| **Project Title** | **Star & Financial AVM** |
| **Duration** | **May 2007 – Dec 2007** |
| **Client** | **WellPoint** |
| **Environment** | **Informatica, Oracle, Teradata** |

**Project Description**

Financial AVM project involves migration of the data from Teradata to Oracle dB on new server and to upgrade all the servers from its existing version to latest version. The scope of this project is confined to ETL Process extracting claims, member, and provider data from CRDW and transforming data based on the business rules and loading data into flat files. The scope of this project is confined to ETL Process extracting claims, member, and provider data from CRDW and transforming data based on the business rules and loading data into flat files.

**Role: Developer**

**Technical Responsibilities**

* Analyzed the business requirements and design specification documents from the design team.
* Designed and developed appropriate ETL mappings to extract and transform data from various sources to meet requirements.
* Created sessions and workflows using workflow manager, configured tasks in the workflows appropriately and extensively used Teradata utilities.
* Debugged the mappings using Informatica debugger and resolved issues using the same.
* Monitored workflows using Informatica workflow monitor and analyzed the session logs for the failed sessions to resolve the issues.

**Organization: Carevoyant Technologies**

|  |  |
| --- | --- |
| **Project Title** | **Patient Clinical Reports** |
| **Duration** | **Sep 2006 – Feb 2007** |
| **Client** | **Carevoyant** |
| **Environment** | **Crystal Reports, Oracle** |

**Project Description**

CareVoyant is a leading developer of integrated Clinical, Financial, Administrative and Business Intelligence health care software solutions based on the latest Microsoft technologies. CareVoyant solutions for Home Care Agencies, Long-term Care Facilities and Outpatient Medical Practices empower providers to streamline workflow, improve quality of care and optimize reimbursement. The scope of this project is to create various patient clinical reports using Crystal Reports like admission info, Guarantor, Patient Plan etc...

**Role: Reports Developer**

**Technical Responsibilities**

* Analyzed the report specification document and design of on demand as well as scheduled Patient clinical reports using Crystal reports designer.
* Developed reports based on the user requirements and report design specification document.
* Tested and debugged the reports.
* Published the reports using crystal reports suite.