Big Data Transformation

Project Name: Project Bluegrass

Program Name: Big Data Transformation

Business Function: Wireless

Location: Atlanta

Region: USA

Project Sponsor: CDO

Estimated Start Date: 10/01/2021

Estimated End Date: 06/01/2022

Date Prepared: 09/01/2021

Prepared by: Rohan Chakravarthi

Contents

1 Business Objective / Problem / Proposed Solution

Business Objective:

TelecomUSA requires a next generation Big Data platform to power the current and future growth of analytics across the company, further enhancing the innovative and reliable operator reputation that TelecomUSA enjoys. This Big Data solution will further enhance innovation, support the growth of self-service BI, drive better decision-making, improve competitive advantage and customer experience. Many new applications such as next best move, churn prediction, hyper personalization, customer journey will be written based on the big data solution. The expected net new revenue generated as a result of this project is estimated at 20M in 2022 with a 100M projection in 5 years.

Problem Statement:

Appliance based solutions drove the last wave of analytics advances but are now outdated and a different approach is required by organisations that need to drive more immediate intelligence and decisions from their precious data. The approach of tightly coupling the software with a bespoke hardware appliance is too inflexible, costly and limiting for the next generation of data analytics.

This initiative requires a complete replacement of the legacy EDW appliance, migrating the current reports, dashboards and BI activity onto the next generation platform to be fully completed, tested and in production by June 2022 when the current appliance is End of Life (EOL) and becomes unsupported.

The next generation EDW will cope with the anticipated workload growth for at least the next 3 years and provide a data warehouse to support both:

- **Operational analytics** - “In the moment” in real-time or near real-time operational reporting activities

- **Advanced analytics** – to provide a better understanding of customers and their behaviours.

Technically TelecomUSA requires new target EDW environments for Development, Test, Pre Production and Production, with an option to provide full Disaster Recovery should that be required at some point. Backup and Recovery is required for both on-premise and cloud based EDW service that delivers close to zero impact on the production environment.

The new target EDW environments must be installed/configured, migrated to and managed, whether on-premise or in the Cloud. This includes migration of the existing legacy EDW Database and Database objects as well as all the data, and all components that read or write to the EDW (e.g. BI tools such as Power BI, Tableau and MicroStrategy).

Proposed Solution:

After full research, Project Bluegrass team is proposing the use performance-beating, innovative, open and highly cost-effective Amazon Redshift technology at the heart of this proposal. Amazon Redshift has the most advanced architecture built specifically for high performance analytics and operational analytics at scale. It uses modern performance processing techniques within its software to deliver linear performance to thousands of users using Petabytes of data. This solution also uses commercially available commodity hardware rather than bespoke hardware that needs to be purchased from an appliance manufacturer. This approach improves flexibility and performance, avoids lock-in and lowers Total Cost of Ownership (TCO).

Move work progressively, and at TelecomUSA's pace, from the on-premise product to the fully-managed cloud service in AWS Cloud. We will work to help provide a seamless transition to the Cloud and do this within timescales that work for TelecomUSA business side to help remove any risks associated with such a strategic initiative.

During this transition, we will work with TelecomUSA wireless business users to determine whether the migration will be workload/application centric or by specific business/scientific teams. We will ensure that performance and availability will be to the same level as the on-premise solution.

2 Project Scope

The project scope includes

- **Production stability** will be delivered not only by the Enterprise class Redshift database, but also leveraging the robustness of Hadoop. An "N+1" process will be adopted to help ensure that should any failure occur the system provides huge levels of high availability

- **Sunset legacy appliance by June 2022.** Our solution delivers a seamless transition from legacy DW to our next generation solution and combines with comprehensive services to ensure a smooth and risk-free path to the end state

- **Seamless transition to the Cloud**, at TelecomUSA's choice of how quickly, how much and when workloads are moved to the Cloud.

- **Replace existing data warehouse** to support business and projected growth for at least the next 3+ years. Our solution replaces the legacy appliance and has been sized to accommodate the current workload and expected growth volume. As a part of the implementation work the new high performance data warehouse (based on Amazon Redshift) will be integrated into the current environment, both for ingestion of data (from DataStage) and analytics and reporting.

- On **Reporting**, our Amazon Redshift solution provides high speed analytics to multiple end users groups and personas. These range from fixed reporting type users who require reports to review SLA/KPIs/Management reports etc, business analysts who need to be able to drill down on dashboards and reports in adhoc ways, all the way through to enabling data scientists to be able to carry out data modelling to get further insight on data. Whether the users are using Tableau, Qlik, MicroStrategy for reporting or 'R' or Scala or Python for data modelling, Big Query performance and flexibility in having all the data optimized all the time means that this is possible

- **Migration** from legacy DW and validation of all reports, dashboards, BI interaction and analytics tools as documented without disruption to the business. Likewise, the data will be migrated without disruption and processing will be run in parallel as part of the migration process. The implementation section of our proposal goes into this in more detail

- **Batch and overnight operations** will easily be carried out in the window required by TelecomUSA business, as we expect to not only meet but comfortably surpass the SLA here. Vector does not require any secondary indexing to be performant. This means the number of write operations will be significantly reduced. This is on top of our previous comparisons against Amazon Redshift which have shown us to be 10x faster.

3 Major Deliverables

Our solution will deliver:

- Highly scalable and performant Enterprise Data Warehouse (EDW)
- Data Warehouse running on powerful, flexible and easily scalable commodity hardware
- Journey to the Cloud – the same underlying technology available on premise and in the Cloud supports an easy route to the Cloud with minimal disruptions

- Migration and implementation services to ensure a swift, efficient and risk-free transition to the new environment using automation tools
- Operations and support of new environment for a fixed cost
- Development, Test and Production environment included
- Fixed and reduced Total Cost of Ownership (TCO) of 30%

4 Requirement Assumptions, Constraints, Dependencies & Risks

Assumptions

Here are the assumptions made while creating this proposal. Any deviation in these may affect the project schedule and cost:

1. Combine production and staging into the same cluster to eliminate need for duplicate staging data load and allowing BI and analytics to be run on entire history
2. Retain all data on-premise according to current security policy
3. No Disaster Recovery environment is to be included in the initial build, in line with the current Amazon Redshift environment.
4. System backups will be performed to existing TelecomUSA SAN.
5. TelecomUSA will provide the environment and required access for the installation, deployment, and testing of automated migration tools and accelerators
6. TelecomUSA will provide Amazon Redshift workloads/queries, associated DDLs, ETL jobs and any other artefacts relevant to the scope of this engagement
7. TelecomUSA will onboard and provide verified security clearance to data migration team
8. TelecomUSA will provide test data for execution and validation of converted workloads
9. Proposal assumes there will be no change in the scripts, schemas, ETL jobs and other artefacts in the scope of conversion during the execution of the project
10. Automated tools and accelerators will be utilised for performing transformation of Amazon Redshift workloads

Constraints

1. Design workshop can be scheduled, only after the SOW is signed
2. Bluegrass team will not be able to support any unsupported feature(s) on any Technology, which may surface during Operations
3. We may have to offer related services from time to time which will result in additional statements of work being raised.

Dependencies

	Dependency Description	Dependency On
1	Hardware setup will be provided by TelecomUSA for the following objective. - Setting parallel environment in development and testing environments for parallel testing - Setting Bi tools in development and testing environments for parallel testing	TelecomUSA IT

	Dependency Description	Dependency On
2	All the artefacts like high level design documents and other information like locations & the actual code for DataStage, MicroStrategy, Amazon Redshift and other BI tools which are in scope will be provided at the beginning of the project to HCL to facilitate our analysis and understanding on existing systems.	TelecomUSA DW team
3	Relevant access privileges and connectivity will be given to project bluegrass team members at the beginning of the project to access databases and applications in scope.	TelecomUSA IT

Risk Assessment
Several risks have been identified which may limit XYZ's ability to successfully deliver the project:

Description (List all major/high level)	Probability of Occurrence	Mitigation Strategy
VPN for offshore connectivity	High	Delay in providing VPN connectivity to project bluegrass offshore location due to any technical or compliance issues from TelecomUSA will impact the project timelines
Availability of stakeholders for quick resolutions	High	Clear identification of stakeholder community, contact, roles and responsibilities
Existing application awareness and documentation	High	Any lack of documentation or application awareness can hamper the project timelines
Database features that may have impact in existing DataStage jobs	High	Data migration and tools upgrade / migration resources from HCL will collaborate in such a way that early detection of such impacts get resolved

5 Out-of-Scope

1. Inflight projects are out of scope
2. IT shall provide prompt access to information concerning the TelecomUSA systems and applications (this information must completely and accurately reflect any procedures or conditions currently in effect). If not provided there will impact to the initial deliverable
3. Bluegrass team will not be responsible/support for any unsupported feature(s) on any Technology, which may surface during Operations.

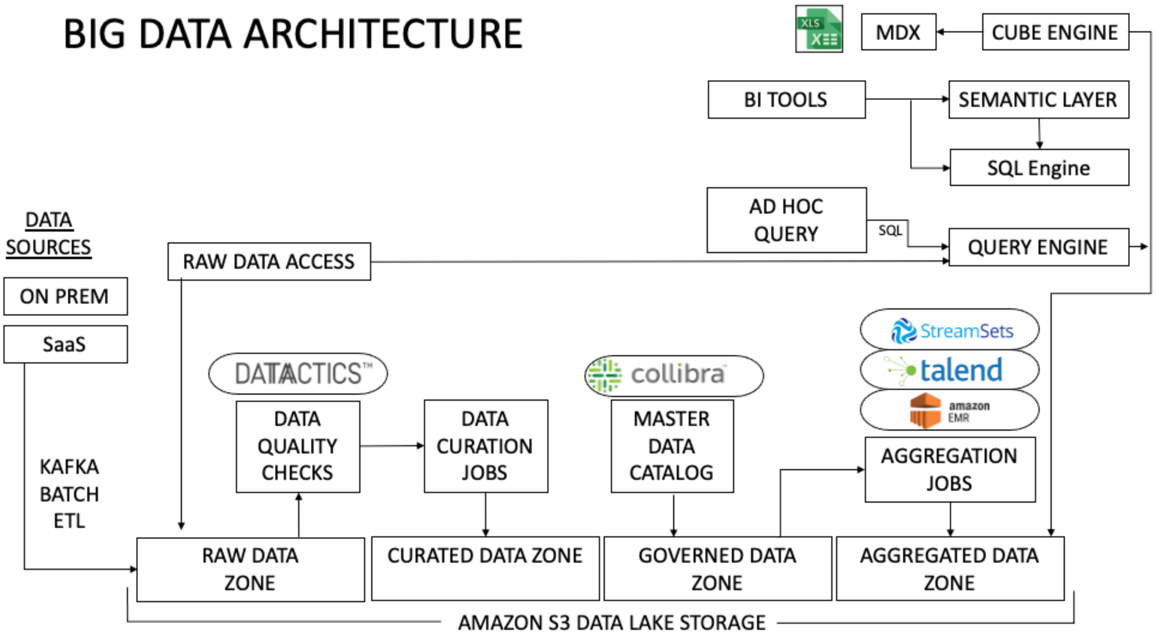
6 Risk if Not Performed

1. If proper user testing of dev, test and pre prod is not performed then there are potential data and reliability risks in production.

2. Disaster Recovery and High Availability region replication tests have to be performed even though the business does not want to fund them at this moment. This can result in a loss of service for an extended period of time in production if not tested prior to going live.

7 Solution Architecture

The solution architecture to support the new data capture and analysis requirements will be defined as a prototype and once confirmed feasibility, will be converted into a production ready environment. The following architectures are recommended for proceeding with this proposal:



Components	Description
Kafka ETL	AWS Kafka based message queue to source records from various data sources
Datactics	Data quality check tool for data cleansing
Collibra	Data governance and master data management tool
Amazon EMR	AWS elastic map reduce to perform aggregration work reading from s3 based storage
Talend	Extract Load Transform tool to load data from source into the SQL engine Redshift
Streamsets	Dataops platform to manage and schedule ELT work
SQL Engine	AWS Redshift SQL engine for big data querying

AWS s3 data lake storage	Fault tolerant resilient data storage solution
Cube engine	Kyligence product to perform fast dashboard reporting via MDX queries from excel

8 Analysis Tools

Total number of end users: 4000

Total number of user licenses: 200

Total number of concurrent licenses: 200 concurrent user sessions

Total server licenses: ULA

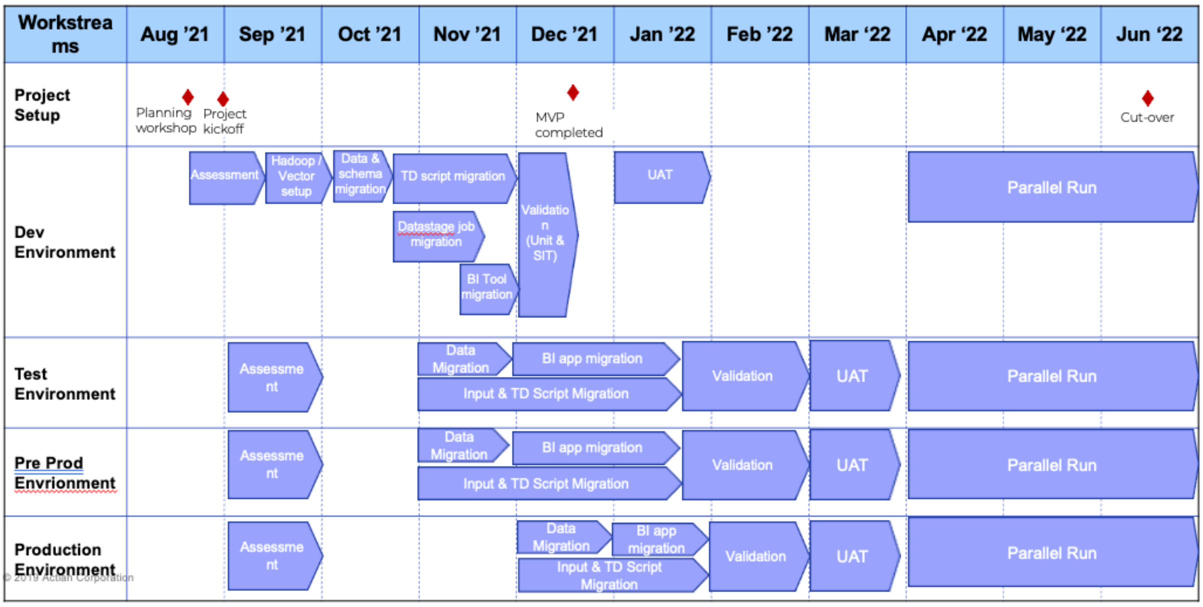
Analysis Tool	**Description**	**Data Focus**	**Cost**
Tableau	BI dashboarding tool for analytics	Customer Journey data	$60/user/month
Kyligence	MDX query tool and plugin to excel	All end user dashboards	$15/user/month

Solution Selected	**Users**	**Cost Per User**	**Total Cost**
Tableau and Kyligence	200 concurrent	$75	$180,000/yr
		Total Budgeted	**$ 180,000/yr**

9 Schedule for Delivery

Here you will develop a week by week schedule to be used to guide the solution progression defining the effort, tasks, and delivery date.

The schedule for the proof of concept only identifies the first 12 weeks of work effort to produce a basic solution that can be tested to ensure the expected outputs can be delivered.



10 Resource Requirements

Significant cross business unit involved. The hours presented are totals for 3-month period for the currently known stakeholders. The stakeholder list will be continuously updated during the project.

Role	Name	Business Unit	Total Effort
Data Engineer	Mary	IT	400 Hours
Data Engineer	Tom	IT	400 Hours
Data Engineer	Bob	IT	400 Hours
Data Scientist	Julie	Data Science	250 Hours
ETL Developer	Sam	IT	600 Hours
ETL Developer	Don	IT	600 Hours
ETL Developer	Kim	IT	600 Hours
Tableau Developer	Chris	Wireless Business	400 Hours
Kyligence Developer	Julia	Wireless Business	400 Hours
Redshift Administrator	Bill	IT	400 Hours
QA Tester	Ronald	Wireless Business	600 Hours
		Total Budgeted	**4950 Hours**

- Royalties, Pricing, Sales and Marketing involvement to be determined during the first (preparation) week. Total of 40 hours expected.

11 Solution Pricing

The work effort within this scope statement will be accomplished at a proposed role based rate as identified below:

Task	Year 1	Year 2	Year 3	Total	Remarks
Operations Services (Maintenance and Support)	550,000	900,000	900,000	2,350,000	Includes HW Support
Amazon Redshift based data warehouse	1,000,000	1,800,000	1,800,000	4,600,000	Includes license cost and supporting software components. TelecomUSA will be responsible for hardware as well as hardware. Also includes Avalanche sandbox for the 1st year
Migration of existing data and corresponding workloads	2,000,000	0	0	2,000,000	Includes cost of tool based and manual conversion of legacy DW scripts. Also includes cost of repointing BI tools and various types of unit, SIT, and end to end testing. TelecomUSA will be responsible for User Acceptance Testing with HCL assistance.
Total	**3,550,000**	**2,700,000**	**2,700,000**	**8,950,000**	

All work will be approved prior to start. All work completed will be accepted by TelecomUSA CDO prior to invoicing. Invoicing will be accomplished on a monthly basis according to approved efforts.

12 Project Justification

Expected Benefits:

- **High performance:** delivered via the Redshift and AWS data lake solution which typically provides 10x performance improvement over Teradata appliances
- **High scalability:** Enabled by using a S3 based cluster, allied to the Redshift engine, running on commodity hardware
- **Production stability:** Delivered not only by the Enterprise class Redshift, but also by adopting an “N+1” replication process to provide high availability
- **Growth:** 3+ years of data growth factored into the sizing of the system

Alternatives Considered:

- Continue with the legacy on prem datawarehouse and pay for the additional extended support cost of $5M and re negotiate contract with legacy vendor

Feasibility Study Results: N/A

13 Conclusion

Implementing this solution for big data enables a new big data environment to achieve the corporate goal of understanding the customer journey and deliver new revenue opportunities by getting the know the customer behavior. Project bluegrass team recommends implementing this new modern cloud basedsolution and migrate away from legacy system.