|  |  |
| --- | --- |
| GENERAL PROJECT INFORMATION | |
| PROJECT NAME | Data Warehousing and Analysis of SEC.gov Financial Form Submissions |
| PROJECT SPONSOR | Professor Ponle Salu |
| PROJECT MANAGER | Sri Venkatesh Subramaniam, Ankit Sharma, Nishit Popat |
| STAKEHOLDERS | Key stakeholders include investors, researchers, regulators, and other stakeholders in the financial sector who rely on SEC data for decision-making. Their primary need is easy access to clean, reliable, and well-organized financial data. |
| EXPECTED START DATE | June 8th, 2023 |
| EXPECTED COMPLETION DATE | August 12, 2023 |
| PROJECT DETAILS | |
| EXECUTIVE SUMMARY | This project aims to address the challenge of extracting, cleaning, and analyzing financial form submissions from the SEC's website. These forms hold rich financial data, under-utilized due to the complexity of handling it. By leveraging Google Big-Query's capabilities, we propose developing a data warehouse solution to efficiently organize and analyze this data, enabling better decision-making for stakeholders in the financial sector. |
| Context | SEC receives a vast amount of financial data from corporations yearly. However, manual analysis of this information is time-consuming, error-prone, and inefficient. Amid rising needs for transparency and accountability in the financial sector, a robust solution for effective data management and analysis is crucial. |
| Problem/Opportunity Statement | The project aims to tackle the problem of data mismanagement and under-utilization of SEC's financial form submissions. Implementing a solution can potentially lead to better transparency, enhanced decision-making capabilities for investors, and greater accountability in the financial sector. |
| EXPECTED BENEFITS | The project targets tangible benefits such as reduced time and effort in data analysis, and intangible ones like enhanced data accessibility, making it easier for stakeholders to make informed decisions. |
| SCOPE | The project covers data extraction, cleaning, transformation, warehousing in Big-Query, and setting up analysis tools. It excludes the development of an entirely new UI for data visualization. |
| Solution Options | 1. Develop a custom solution from scratch (hiring a SQL dev and building a local in-premise server).  2. Utilize Google Big-Query for a comprehensive data warehousing and analysis solution. |
| Assessment of Options | |  |  |  |  | | --- | --- | --- | --- | | Criteria | Scale | Custom Data Warehouse | Google Cloud Platform(GCP) - Big Query | | Cost | 1-Extremely expensive: $16000 - $12000  2-High: $12000 - $9000  3-Moderate: $9000 - $6000  4-Low: $6000 - $3000  5-Small cost: <$3000 | 15,000 | 600 | | Time to Implementation | 1-Long: > 1 year  2-Moderately long: 9-12 months  3-Acceptable: 6 months  4-Good: 3 months  5-Fast: < 3 months | 6 Months | 2 Months, 2 Weeks | | Technical expertise required | 1-High: Extensive expertise  2-Moderate: Intermediate expertise  3-Acceptable: Basic expertise  4-Low: Minimal expertise  5-None: No expertise required | Very High,  requires extensive technical expertise to handle system integration, data management, and configuration. It may involve complex software setup and customization. And System Administration | Moderate knowledge of GCP and SQL | | Return On Investment | 1-Low: < 5%  2-Moderate: 5-10%  3-Acceptable: 10-15%  4-Good: 15-20%  5-Excellent: > 20% | Estimated 5% | Estimated 100% | | Risks | 1-High: Multiple high risks  2-Moderate: Moderate risks  3-Acceptable: Some risks  4-Low: Minimal risks  5-None: No risks | High.  There are high risks associated with implementing a new Data Warehouse, including high implementation costs, potential delays, lack of scalability, Manual system administration, and security vulnerabilities. | Very Low.  Minimal training cost,  GCP cost,  Expertise and guidance available with Project Sponsor | | Scalability | 1-Limited: Limited scalability  2-Moderate: Moderate scalability  3-Acceptable: Acceptable scalability  4-Good: Good scalability  5-Excellent: Excellent scalability | Limited Scalability  Storage and Hard-drive partitioning and CPU constraints | Excellent  GCP Provides excellent scalability options with low cost for PetaBytes of Data | | Maintainability | 1-High: High maintenance requirements  2-Moderate: Moderate maintenance requirements  3-Acceptable: Acceptable maintenance requirements  4-Low: Low maintenance requirements  5-None: No maintenance requirements | Highly challenging to administer and manage the warehouse server. And to ensure optimized performance. | Low Maintenance with GCP services as GCP provides automated maintenance services. | | Security | 1-Low: Low security measures  2-Moderate: Moderate security measures  3-Acceptable: Acceptable security measures  4-Good: Good security measures  5-Excellent: Excellent security measures | Good, The warehouse system can offer good security with user access an credential management. | Excellent security as the GCP servers are hosted in house with Google Computers. | | Score Total: |  | 13 | **35** |   Custom Data Warehouse: This approach allows for the highest degree of customization. However, it may require considerable time and financial resources, and the technical challenge of handling big data could potentially lead to setbacks. Due to the setup of local machine and server.        Google BigQuery Solution (Google Cloud Platform): BigQuery offers scalable, efficient data management and analysis. Despite having less customization than a bespoke solution, it is cost-effective, faster to implement, and leverages Google's robust cloud infrastructure.    Given these assessments, we recommend the Google BigQuery solution due to its cost-effectiveness, reliability, and the speed of implementation. |
| Detailed List of Deliverables | 1. A fully functional data pipeline of SEC.gov public dataset into Big-Query. 2. Data Warehouse deliverables:    1. A transformed, custom tables and views suitable for analysis.    2. Implementation of custom procedures and views for quick and robust access of relevant data. 3. Document Deliverables:    1. Quick access guide    2. Documentation of system and query usage. |

**Project Decision-Making Process Flow Chart**

A diagram of a problem

Description automatically generated with low confidence

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Timeline  and Milestones | |  |  |  | | --- | --- | --- | | PHASE | DEADLINE | DESCRIPTION | | PHASE 1 | June 8 | Setup of data extraction and loading into BigQuery. | | PHASE 2 | June 15 | Data cleaning and transformation. | | PHASE 3 | July 3 | Team Training | | PHASE 4 | July 4 | Data warehousing setup in BigQuery. | | PHASE 5 | July 29 | Setup of data analysis and reporting tools.  Documentation | | PHASE 6 | August 9 | Testing and final implementation. | |
| Risks and  Dependencies | 1. Data Quality: If the data from SEC.gov is not of high quality, it could impact the results and introduce complexity in data transformation. 2. Technical Risks: There could be technical challenges or setbacks during development due to limited knowledge and expertise. 3. Change in SEC Data Policies: Changes in data submission policies by SEC such as data management rules could affect the project with data retrieval and modeling |
| Assumptions | 1. The data from SEC.gov will continue to be open for retrieval. 2. The data from SEC.gov will be regularly updated quarterly. 3. The stakeholders have access to Google BigQuery and relevant data visualization tools. |
| Recommendation | We recommend implementing the Data Warehouse solution in BigQuery provided by Google Cloud Platform, considering its efficiency, reliability, cost-effectiveness, and speed. This solution will enhance decision-making capabilities, leading to transparency and accountability in the financial and capital markets sector. |

|  |  |
| --- | --- |
| Risks and  Dependencies | * 1. Data Quality: If the data from SEC.gov is not of high quality, it could impact the results and introduce complexity in data transformation.   2. Technical Risks: There could be technical challenges or setbacks during development due to limited knowledge and expertise.   3. Change in SEC Data Policies: Changes in data submission policies by SEC such as data management rules could affect the project with data retrieval and modeling |
| Assumptions | 1. The data from SEC.gov will continue to be open for retrieval. 2. The data from SEC.gov will be regularly updated quarterly. 3. The stakeholders have access to Google BigQuery and relevant data visualization tools. |
| Recommendation | We recommend implementing the Data Warehouse solution in BigQuery provided by Google Cloud Platform, considering its efficiency, reliability, cost-effectiveness, and speed. This solution will enhance decision-making capabilities, leading to transparency and accountability in the financial and capital markets sector. |
| Next Steps | 1. Obtain Approval from Project Sponsor 2. Assemble Team 3. Kick off meeting and scheduling process 4. Begin Development 5. Regular Check-ins: Maintain regular project updates to keep all stakeholders informed. |