# Enterprise Data Analytics Strategy

**Purpose**: This document describes the vision of the City of San Jose for an Enterprise Data Analytics Platform **(EDAP)** and an implementation roadmap for its development.

**Business Objectives**: Provide a single platform for all required data assets to be available for city data analysts, data scientists, business managers, executives, and policy makers to be able to conduct holistic, complete, and accurate data analytics for serving the residents and businesses of the City of San Jose in an equitable, effective, efficient, and transparent way**.** In addition to fostering and supporting data-driven decision making and analytics by City government managers and policy makers, the Enterprise Data Analytics Platform will also support the publication of pertinent data on the City's Open Data Portal. This will eliminate standalone processes that only cover open data and will lead to a more efficient and broader scope for the City's open data assets.

**Strategy**: **Develop an Enterprise Cloud Data Lake/Warehouse and Processing Framework**

Building an Enterprise Cloud Data Lake/Warehouse and the Processing Framework for the City of San Jose to support the business objectives and provide significant benefits and achieve cost savings:

1. Capability to ingest any type of data into the Enterprise Cloud Data Lake/Warehouse using automated pipelines.
2. Data consumers be able to access data assets from different business domains for analytics workloads, machine learning models, and business intelligence reporting, in one central location. Data becomes readily, reliably, and cost effectively available to business users who can extract insight through data analytics applications for better serving the business requirements of the city.
3. Data assets in the Data Lake/Warehouse will be managed consistently and reliably through an open governance framework and the maintenance of a central data catalog.
4. A consolidated metadata framework will be put in place to ensure schema, attribute, and semantic consistency across all data assets available in the Data Lake/Warehouse.
5. Privacy protection and secure access for authorized users will be enforced for all data assets in the Data Lake/Warehouse.

Standardized practices in supporting and maintaining the Enterprise Data Lake/Warehouse will result in better, more valuable, more efficient data analytics services, economies of scale, and cost savings.

**Governance**: A data analytics platform consumes data from various operational data sources. As such, the governance of the enterprise data analytics platform requires alignment with the City of San Jose's enterprise data architecture governing body. An Enterprise Data Analytics Governance Group will be established to include representation by key stakeholders in developing and supporting EDAP. Data governance is the convergence of data management, data quality, data policies, and risk management surrounding the handling of data in the organization. Successful and effective data governance requires deploying a Data Catalog software platform.

**Implementation Roadmap**:

1. **Align Data Analytics Strategy Goals with City's Strategic Service Framework**: City services are constituent-centric. It is important to build the Enterprise Data Analytics Platform to support analytics for a holistic view of the various services provided to the client.
2. **Identify High Value Data Assets (Data Catalog)**: In collaboration with city elected officials, City Manager, department executives, community groups, and researchers in local colleges and universities to identify service areas that could be improved significantly using data analytics.
3. **Develop Objectives and Key Results (OKR)**: A set of concrete Service domain Business Objectives and related Key Results will be developed to be used as strategic drivers for building insightful data analytics models.
4. **Establish a Data Sciences Center of Competency (or Excellence)**: A City of San Jose Data Sciences Center of Excellence will be formed to provide leadership, direction, implementation guidelines, and share knowledge for consistent utilization of data analytics, and application of best practices and objective methodologies in all service domains using the new Enterprise Analytics Platform.
5. **Develop the Architecture for the Data Analytics Technology Platform**
6. **Implement a Pilot Project**: The City of San Jose Department of Transportation has significant investments in Power BI and related Azure services. We will implement the Enterprise Analytics Platform Pilot Project for …
7. **Evaluate the Pilot Implementation**
8. **Implement the Enterprise Data Analytics Platform**

---

1. **City's Strategic Service Framework:** The ultimate goal of data analytics is to be an effective tool for business decision makers to conduct city policy development and service operations in the most effective, efficient, and transparent way. This goal can be achieved through a commitment and practice of data-driven decision making. The road to advanced analytical maturity is a process that begins with building the infrastructure and developing Descriptive analytics for the core service domains in the city. This path can then be expanded to higher levels of analytics including Predictive, and ultimately, Prescriptive analytics. The Department of Transportation (DOT) has been selected for the Pilot implementation. DOT "plans, develops, operates, and maintains transportation facilities, services, and related systems that contribute to the livability and economic health of the city".

2. **High Value Data Assets (Data Catalog)**: Managing data assets begins with identifying and tracking what data repositories we have. As a first step, we will deploy a **Data Governance software platform** (e.g., **Apache Atlas**) to catalog our existing data assets and use metadata management to track data ownership, data quality, lineage, data access policies, and a consistent glossary of terms. It should be noted that the Enterprise Data Lake shall include GeoSpatial data and the analytics framework will include components and tools for GeoAnalytics as well.

   In the initial phase, we have identified the following data assets maintained by the San Jose Department of Transportation that will be added to the Enterprise Data Lakehouse:

| Division | Source System | Dataset Name | Existing Analytics Applications |
|---|---|---|---|
| Infrastructure Maintenance | | | |
| Transportation Safety, Operations, and Parking | | | |
| Transportation Planning and Project Delivery | | | |

3. **OKRs**: The city strives to provide quality services to its residents and constituents in the most equitable, transparent, and efficient and cost-effective way. Data assets are the key components in increasing consistency and confidence in decision making. Data assets can also be used to develop a quantifiable system of measures to track performance of city agencies in delivering services to the public.

   It is important to adopt a methodology for developing business metrics to ensure objectivity, consistency, and validity of the metrics used. This effort is not driven by technology but rather by engaging the right business stakeholders and implementing measures derived from "Evidence-based Practice" and scientific methods. This approach also focuses on optimizing the quality and quantity of the measures developed. In the ideal balance, a limited and highly practical number of metrics shall be deployed to avoid overloading business decision makers with too much information.

   Data analytics presents business insights using metrics that provide quantitative measures of the state of business operations and their outcomes. Measures can be grouped into four categories (using the Gartner Data Analytics Maturity Model):

   a. **Descriptive** – describing what "happened". These measures are primarily focused on Inputs and Outputs of a service operation. They report counts, averages, or totals for operations that serve the public.

b. **Diagnostics** – why it "happened".
c. **Predictive** – what will "happen".
d. **Prescriptive** – how can we make it "happen".

**Service Performance Metrics Framework**: It is critical that we measure the right metrics accurately and consistently. We need to establish an enterprise framework for defining and calculating metrics that are essential to the city's operations. The benefits of this framework include:

a. Consistency: ensures that metrics are defined and calculated consistently across all service domains.
b. Flexibility: allows the city to continue to add new measures and improve existing ones as the business requirements change.
c. Transparency: makes it easy to understand how metrics are calculated and therefore increases trust in the data and the metrics.

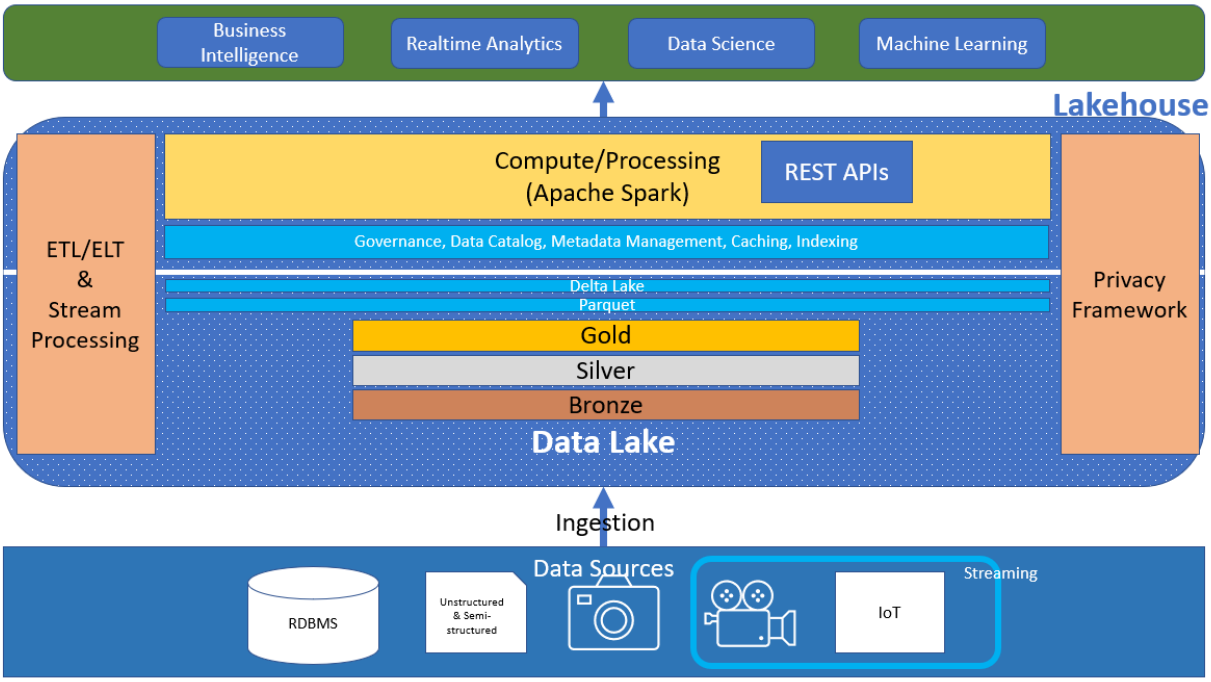
| Examples | | |
|---|---|---|
| **Division** | **Metrics** | **Data Source** |
| Transportation Planning and Project Delivery | Number of drive-alone trips | |
| Transportation Planning and Project Delivery | Vehicle miles traveled (VMT) per capita | |
| | | |

4. **Data Sciences Center of Competency**: This strategy calls for establishing an Enterprise Data Sciences Center of Competency comprised of city data management and analytics staff (could be augmented with contract resources as needed) who will develop and support the Enterprise Data Analytics Platform. It's important to have city staff fill this role to ensure maintaining institutional knowledge and continuity and long-term sustainability of the city's data analytics program.

5. **Architecture**: Successful implementation of an enterprise data analytics strategy requires designing an open and well-thought-out architecture and selecting the right technologies that are based on industry standards and proven methodologies.

The City of San Jose Data Analytics Strategy implements a **multi-layer architecture** to ensure technology openness, compliance with industry standards, taking full advantage of best practices and the latest advancements in data management and analytics technologies, and the flexibility to upgrade and implement future innovations as the technology continues to evolve.

**High Level Architecture - Overview**




**Multi-Layer Architecture:**

i. Deleting and Updating Data to **Comply with Privacy Regulations**: Privacy regulations such as the General Data Protection Regulation and California Consumer Privacy Act require the need for record-level deletes on the data lake for "right to be forgotten" or to store changes to consent on how the constituent's data can be used. The ability to perform record-level deletes and updates is a fundamental requirement and therefore is built into this architecture. For compliance regulations, the platform shall support functionality for hard deletes (deleting records from the table and physically removing them from the data lake).

ii. Data Design Pattern: We shall use **Medallion Architecture** as the data design pattern to logically organize data in the Lakehouse, with the goal of incrementally and progressively improving the structure and quality of data as it flows through each layer of the architecture (from Bronze to Silver to Gold).

iii. Data Storage: We shall use a **Cloud Data Storage** solution.

iv. Data Storage Format: We shall use **Apache Parquet** format to store data. Parquet is an open-source file format that stores data in columnar format (as opposed to row format) that can be

processed much more efficiently and cost-effectively than other formats, making it an ideal file format for big data, analytics, and data lake storage.

v. Data Lake Table Format: We shall use **Delta Lakehouse** for managing the raw data stored to provide for record-level operations, ACID transactions, time travel, rollbacks, and caching. Table formats provide additional database-like functionality for data lake files.

vi. Change Data Capture (CDC): Data stored in the Lakehouse must be continuously kept up to date to mirror data in the source systems. Change Data Capture is a software process used to replicate actions performed against operational databases for use in downstream analytics applications. There are different implementation models for CDC, such as Pull-based CDC, Push-based CDC, and Log-based CDC. We will choose a CDC model that would best fit the requirements of our architecture as part of the Pilot Project implementation and evaluation.

vii. Extract, Transform, Load (ETL, ELT) and Streaming Data: We shall use an industry leader product to perform read and write operations on Delta Lake tables. For streaming data, we shall use Apache Spark (Structured Streaming) and Apache Kafka to process the data and store them in the data lake.

viii. Data Catalog: An **Enterprise Data Catalog** shall be developed and maintained to effectively govern and manage data assets stored in the data lake. The data catalog shall support the functionalities of a complete data management lifecycle, from data discovery and search to identifying and designating sensitive data that need to be protected, masked, or anonymized.

ix. Processing Engine: We shall use the open-source **Apache Spark** framework to accelerate time to insight across all data assets stored in the enterprise data lake. Apache Spark can execute large-scale data transformations and analyses, run advanced machine learning algorithms, and graph processing applications. It is a multi-language engine for executing data engineering, data science, and machine learning.

x. REST APIs: All datasets in the Enterprise Data Analytics Platform shall be made available to authorized users under the "**Data as a Service (DaaS)**" model. We shall provide REST API access for services available on the platform. We shall use the **Apache Spark REST API and the Apache Livy**. Apache Livy is a service that enables easy interaction with a Spark cluster over a REST interface.

xi. Analytics Dashboards and Machine Learning: We shall develop data analytics dashboards to provide insights and decision support to the city's business stakeholders and support advanced data analytics services including machine learning.

## 6. Pilot Project: A Pilot Project using datasets from the Department of Transportation will be implemented to pave the way for the full deployment of the Enterprise Analytics Platform. This includes identifying the deployment path (i.e., Data Storage, ETL/ELT, Processing, one or more Dashboards (Descriptive Analytics), and a demonstration of the use of advanced analytics (Predictive Analytics).

The Pilot will also include the implementation of a Cloud-based Data Catalog tool to cover data assets that will be added to the Data Lakehouse. The Catalog will provide maps of the data landscape with automated data discovery, sensitive data classification and end-to-end data lineage.