System Blueprint to Support Data Science, Machine Learning, and AI

Acquire → Store → Cleanse → Visualize → Analyze

Extract - Transform - Load

1 DATA SOURCES

Internal

- Excel, CSV, Images
- Email & Sharepoint
- Project/Program database
- Operating data
- AMI / Meter data
- Sensor data
- Databases and data warehouses

External

- Government & public data
- Weather data
- Vendor data
- Contractor data
- Consumer data
- Other 3rd Party data

2 DELIVERY & EXTRACTION

- Manual Connectors
- Database Connectors
- Cloud
- FTP
- Pi2Pi Connector
- APIs
- Streaming

Staging Environment

3 DATA CLEANSING

- Anonymize
- Assess
- Normalize
- De-dupe
- Cleanse
- QA
- Malware

4 VIRTUAL DATA LAKE

Program Research Data

Department database

Central Storage Managed by IT

Database Clusters

Hadoop

PI Systems

Cloud Storage (Public, Private, Hybrid)

5 DATA SCIENCE TOOLS

6 COMPUTE LAYER & VIRTUAL MACHINES

7 DATA SCIENCE WORKSPACE

Visualizations

.pdf

Analyses

- Descriptive
- Diagnostic
- Prognostic
- Predictive
- Text Analytics
- Optimization
- Machine Learning / AI

8 METADATA & DATA MANAGEMENT

9 SECURITY & GOVERNANCE (including CII[1], PII[2], user access, permissions)

EPRI ELECTRIC POWER RESEARCH INSTITUTE

Part of 3002015801

1 CII - Critical Information Infrastructure
2 PII - Personally Identifiable information

System Blueprint to Support Data Science, Machine Learning, and AI

1 DATA SOURCES

New data sources are becoming available daily from public sources such as government agencies, device manufacturers, and internet users through crowd sourcing. This Blueprint lists some examples of internal and external data sources, but the list is much longer and constantly changing.

2 DELIVERY & EXTRACTION

This component involves data science tools to extract and consolidate data from primary databases in bulk or batch. The tools offer an efficient and systematic way to pull in volumes of data. Typically, the data travels to a staging environment for virus and malware screening, before moving into storage.

3 DATA CLEANSING

Cleansing is a critical first step before conducting any data mining or advanced analytics with datasets. Some of the activities include anonymizing data (e.g., removing confidential, identifiable customer information), normalizing data into the same unit of measure or same time of day, removing duplicates, and understanding the magnitude and significance of missing values.

4 VIRTUAL DATA LAKE

This Blueprint expects that all data owners and managers continue to store their data "in place" with no change to its current location. The virtual data lake indexes all the datasets and makes them searchable and available for use by others within the company. Permission for use is determined and granted by the data owner.

5 DATA SCIENCE TOOLS

The tools include open source products like Python, RapidMiner, as well as proprietary platforms such as SAS, Oracle Data Mining, and IBM SPSS Modeler. These tools help data scientists discover predictive information that will help them create analytic models and successfully get to insights.

6 COMPUTE LAYER & VIRTUAL MACHINES

The compute layer refers to the data processing power required to churn through volumes of data for visualization and advanced analytics. Processor-intensive work no longer requires physical machines or super computers. Today, companies can scale up with virtual machines (cloud-based) to meet their changing needs for processing power.

7 DATA SCIENCE WORKSPACE

This is a virtual sandbox for creating data visualizations and developing analytic models. For people working in data science, visualize and analyze are the most enjoyable stages of the lifecycle because they lead to new insights.

8 METADATA & DATA MANAGEMENT

Metadata refers to capturing descriptive information about datasets such as data source, data owner, and timeframe that data was collected. Metadata is critical to enable data sharing as it provides the information to create an index to make datasets searchable.

Data management is the organization of datasets and administration of permissions to review, edit, and use data in the virtual data lake.

9 SECURITY & GOVERNANCE

Processes and rules for governance are needed to screen, evaluate, and index datasets before they are stored in the virtual data lake. Governance helps to ensure data lake contents remain relevant and useful. Security protocols are needed to design the user permissions for who can read, edit, and use the data.