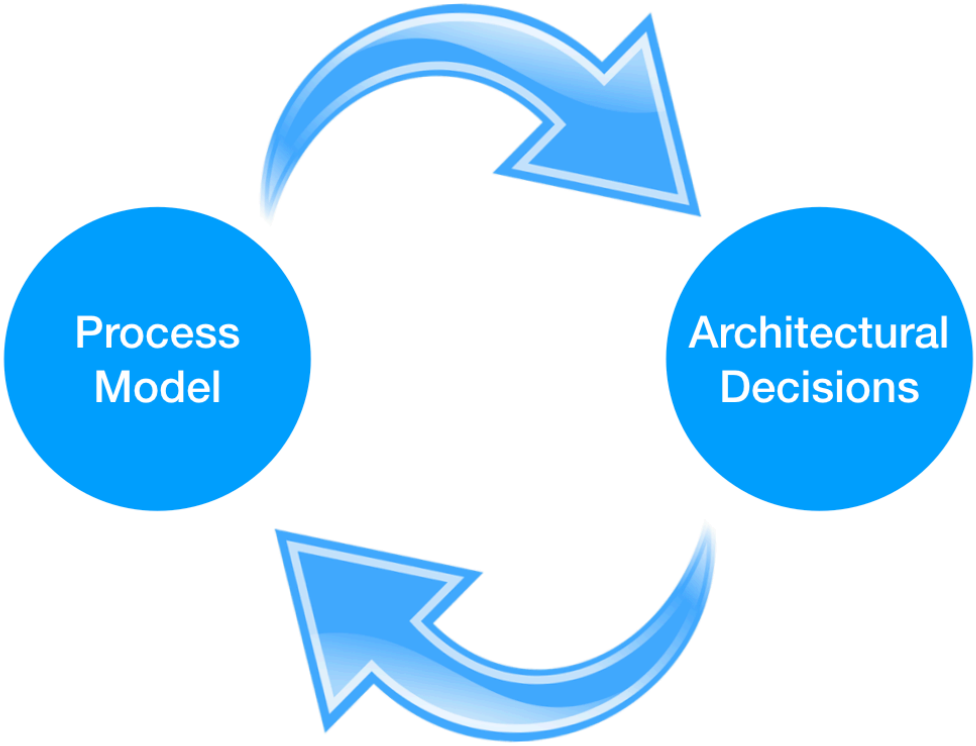
The Lightweight IBM Cloud Garage Method for Data Science

The Lightweight IBM Cloud Garage Method for Data Science includes a process model and an architectural decisions guide to map individual technology components to the reference architecture and guidelines for deployment considerations. The Lightweight IBM Cloud Garage Method for Data Science does NOT include any requirement engineering / design thinking tasks. Since it is hard to initially define the architecture of a project this method supports architectural changes during the process model.

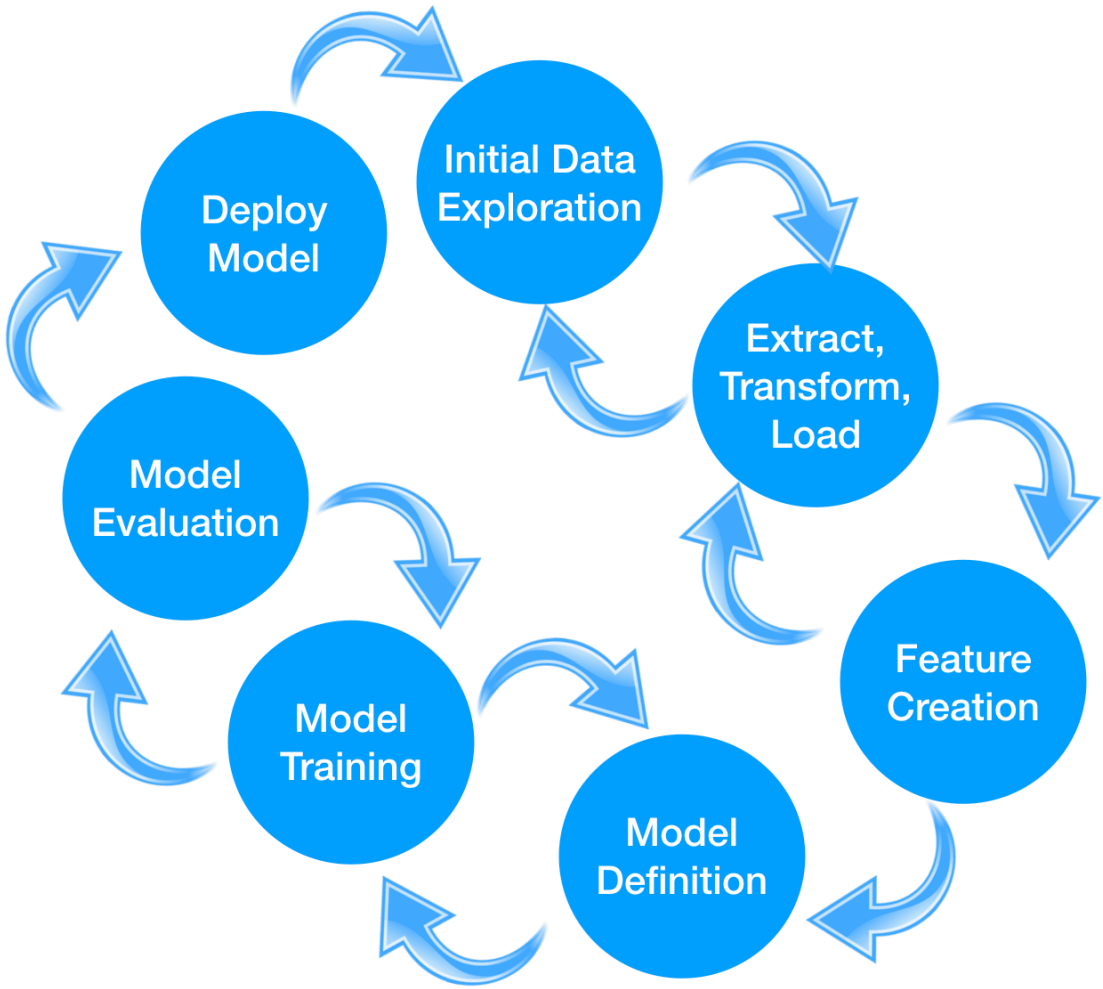


The Process Model influences Architectural Decisions iteratively during the project. Source: IBM Corporation

Table of Contents

1 The Lightweight IBM Cloud Garage Method for Data Science Process Model

In this section, I'll introduce this lightweight process model.



The Lightweight IBM Cloud Garage Method for Data Science Process Model. Source: IBM Corporation

The first thing you should notice is its similarity to the process models we have introduced in my last article TODO link

In addition, there are no design tasks since this method is especially useful for projects where the business expectations are already set.

The last thing you might notice is the increased granularity in the individual tasks.

The reason for this is reuse – every task has a clear purpose and a defined work product (e.g. a jupyter notebook, a script or a docker container hosting a scoring or training endpoint, depending on the architectural decisions made).

In the following, the tasks are explained.

1.1 Initial Data Exploration

1.1.1 Task Guidance

1.1.1.1 Purpose / Objectives

This task is crucial for understanding your data. Data Quality is the most important driver for success in any Data Science project. So, this task gives you the opportunity to address Data Quality just from the beginning, which includes going back to the data owners and asking them for better quality data, if applicable.

1.1.1.2 Reference Materials

https://www.coursera.org/learn/data-science-methodology Module 2 - Data Understanding

1.1.2 Tool Guidance

- IBM Watson Studio jupyther notebooks, scikit-learn, pandas, matplotlib
- IBM Watson Studio jupyther notebooks, Apache Spark, pixiedust
- IBM Watson Studio - IBM SPSS Modeler – Data Audit
- IBM SPSS Modeler Standalone – Data Audit
- IBM Information Server – Quality Stage

1.1.3 Asset Naming Convention

[project_name].data_exp.<technology>.<version>.<extension>

Legend: [] mandatory, <> optional

1.2 Extract, Transform, Load (ETL)

1.2.1 Task Guidance

1.2.1.1 Purpose / Objectives

This task is an important step in transforming the data from the source system into a shape suitable for analytics. In traditional data warehousing this incudes accessing the OLTP system's databases, transforming from a highly normalized data model into a star of snowflake scheme, and finally storing data to a data warehouse. In data science project this step is usually much simpler. Data arrives already in an exported format (e.g. JSON or CSV).

But sometimes de-normalization must be done as well. Finally, the result usually ends up in a bulk storage like Cloud Object Store.

1.2.1.2 Reference Materials

https://www.coursera.org/learn/data-science-methodology Module 2 - Data Preparation

1.2.2 Tool Guidance

- IBM Watson Studio jupyther notebooks, scikit-learn, pandas
- IBM Watson Studio jupyther notebooks, Apache Spark, Apache Spark SQL
- IBM Watson Studio - Data Refinery
- IBM Information Server – Data Stage

1.2.3 Asset Naming Convention

[project_name].etl.<technology>.<version>.<extension>

1.3 Feature Creation

1.3.1 Task Guidance

1.3.1.1 Purpose / Objectives

This task transforms input columns of various relations into additional columns to improve model performance. A subset of those features can be created in an initial task (e.g. one-hot encoding of categorical variables, normalization of numerical variables). Some others require business understanding or multiple iterations to be considered. This task is one of those benefiting the most from the highly iterative nature of this method.

1.3.1.2 Reference Materials

https://www.coursera.org/learn/data-science-methodology Module 2 - Data Preparation

1.3.2 Tool Guidance

- IBM Watson Studio jupyther notebooks, scikit-learn, pandas
- IBM Watson Studio jupyther notebooks, Apache Spark, Apache Spark SQL
- IBM Information Server – Data Stage

1.3.3 Asset Naming Convention

[project_name].feature_eng.<technology>.<version>.<extension>

1.4 Model Definition

1.4.1 Task Guidance

1.4.1.1 Purpose / Objectives

This task defines the machine learning or deep learning model. Since this is a highly iterative method various iterations within this task or including up- and downstream tasks are possible. It is highly recommended to start with simple models first for baseline creation, once those models are evaluated.

1.4.1.2 Reference Materials

https://www.coursera.org/learn/data-science-methodology Module 2 - Modeling

1.4.2 Tool Guidance

- IBM Watson Studio jupyther notebooks, scikit-learn, pandas
- IBM Watson Studio jupyther notebooks, Apache Spark, Apache SparkML, Apache SystemML
- IBM Watson Studio - IBM SPSS Modeler
- IBM SPSS Modeler Standalone

1.4.3 Asset Naming Convention

[project_name].model_def.<technology>.<version>.<extension>

1.5 Model Training

1.5.1 Task Guidance

1.5.1.1 Purpose / Objectives

In this task, the model is trained. This task is set apart from model definition and evaluation for various reasons. First, training is a computationally intense task which might be scaled on computer clusters or GPUs. Therefore, an architectural cut is sometimes unavoidable. (E.g. model definition happens in Keras, but training happens on a Keras model export using Apache SystemML on top of Apache Spark running on a GPU cluster). In the case of hyper parameter tuning and hyper parameter space exploration the downstream task “Model Evaluation” can be part of this asset.

1.5.1.2 Reference Materials

https://www.coursera.org/learn/data-science-methodology Module 2 - Modeling

1.5.2 Tool Guidance

- IBM Watson Studio jupyther notebooks, scikit-learn, pandas
- IBM Watson Studio jupyther notebooks, Apache Spark, Apache SparkML, Apache SystemML

- IBM Watson Studio - IBM SPSS Modeler
- IBM SPSS Modeler Standalone

1.5.3 Asset Naming Convention

[project_name].model_train.<technology>.<version>.<extension>

1.6 Model Evaluation

1.6.1 Task Guidance

1.6.1.1 Purpose / Objectives

In this task, the model is performance is evaluated. Given the nature of the task different metrics must be applied. E.g. categorical-cross entropy for a multi-class classification problem. It is important to divide the data set into training, test and validation (if cross-validation isn't used) and that performance of different feature engineering, model definition and training parameters are kept track of.

1.6.1.2 Reference Materials

https://www.coursera.org/learn/data-science-methodology Module 2 - Evaluation

1.6.2 Tool Guidance

- IBM Watson Studio jupyther notebooks, scikit-learn, pandas
- IBM Watson Studio jupyther notebooks, Apache Spark, Apache SparkML, Apache SystemML
- IBM Watson Studio - IBM SPSS Modeler
- IBM SPSS Modeler Standalone

1.6.3 Asset Naming Convention

[project_name].model_evaluate.<technology>.<version>.<extension>

1.7 Model Deployment

1.7.1 Task Guidance

1.7.1.1 Purpose / Objectives

In this task, the model is deployed. This task heavily depends on the use case. Especially, on the stakeholder's expectation on consuming the data product. So, valid ways of deployment include:

- an interactive jupyter notebook
- an export of an already run, static jupyter notebook, some sort of report
- a REST endpoint allowing scoring (and training) of the model (e.g. backed by a docker container running on Kubernetes)
- a full-fledged web- or mobile application

1.7.1.2 Reference Materials

https://www.coursera.org/learn/data-science-methodology Module 3 - Deployment

1.7.2 Tool Guidance

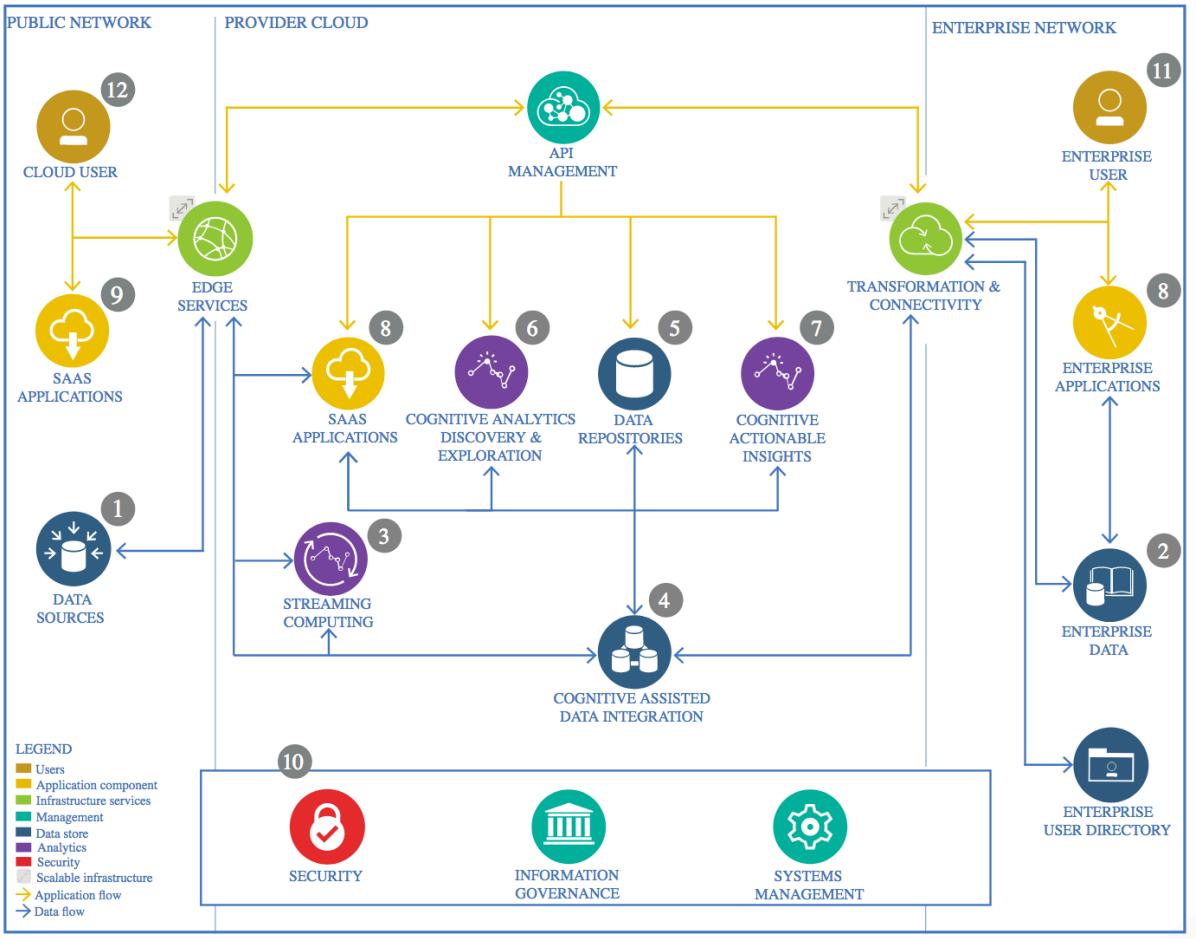
- IBM Watson Studio jupyther notebooks, scikit-learn, pandas
- IBM Watson Studio jupyther notebooks, Apache Spark, Apache SparkML, Apache SystemML
- IBM Watson Studio - IBM SPSS Modeler
- IBM SPSS Modeler Standalone
- IBM MAX (Model Asset Exchange)
- IBM FfDL (Fabric for DeepLearning)
- IBM Watson Machine Learning
- IBM Watson DeepLearning as a Service

1.7.3 Asset Naming Convention

[project_name].model_deployment.<technology>.<version>.<extension>

2 Architectural Decisions Guide

The IBM Data and Analytics Reference Architecture defines possible components on an abstract level. Goal of this section is to choose from the required components and assign real and concrete architectural components to it. Again, this method is highly iterative, so finding during following of the process model can always result in changed to architectural decisions. Please remember, there are never wrong architectural decisions, since they are made in an informed way taking all the knowledge available at a certain point in time into account. The only thing is to document why a decision was made. Below, the IBM Reference Architecture for Cloud Analytics is illustrated. Following this illustration, concrete guidelines are given for each component what technology should be chosen and if the component needs to be included at all.



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

2.1 Data Source

2.1.1 Definition

An internal or external data source which includes Relational Databases, Web Pages, CSV, JSON or Text Files, Video and Audio data

2.1.2 Architectural Decision Guidelines

In fact, on the data source there is not so much to decide since in most of the cases the type and structure of a data source is already defined and controlled by stakeholders. In case there is some control over the process, the following principles should be considered:

- How does the delivery point look like?
  Enterprise data mostly lies in relational databases serving OLTP systems. It is usually bad practice to access those systems directly – even in read-only mode – since ETL processes are running SQL queries against those systems which can bring down performance. One exception is IBM DB2 Workload Manager since it allows OLAP and ETL workload to run in parallel with OLTP workload without performance degradation of OLTP queries using intelligent scheduling and prioritizing

mechanisms. More on IBM DB2 Workload Manager can be found here: https://www.ibm.com/support/knowledgecenter/en/SSEPGG_10.1.0/com.ibm.db2.luw.admin.wlm.doc/com.ibm.db2.luw.admin.wlm.doc-gentopic1.html

- Does real-time data need to be considered?
  Real-time data comes in various shapes and delivery methods. The most prominent include MQTT telemetry and sensor data (e.g. coming from the IBM Watson IoT Platform), a simple REST-HTTP endpoint which need to be polled or a TCP or UDP socket. If no down-stream real-time processing is required those data can be staged (e.g. using Cloud ObjectStore). If down-stream real-time processing is necessary, please read the flowing chapter on "Streaming analytics".

2.2 Enterprise Data

2.2.1 Definition

Cloud based solutions tend to extend the enterprise data model. Therefore, it might be necessary to continuously transfer subsets of enterprise data to the cloud or access those in real-time through a VPN API gateway.

2.2.2 Architectural Decision Guidelines

2.2.2.1 Batch or real-time access

Moving enterprise data to the cloud is always costly. Therefore, it should be considered only if necessary. E.g. if user data is handled in the cloud – is it sufficient to store an anonymized primary key. If transfer of enterprise data to the cloud is unavoidable, privacy concerns and regulations must be addressed. Once legal aspects are taken care of there exist two ways of access. Batch wise sync from enterprise data center to the cloud or real-time access to subsets of data using VPN and an API gateway. This depends on the up-to-dateness requirements.

2.2.2.2 Technology Guidelines

2.2.2.2.1 Secure Gateway

Secure gateway allows cloud applications to access specified hosts and ports in a private data center though an outbound connection. Therefore, no external inbound access is required.

More Information:
https://console.bluemix.net/catalog/services/secure-gateway

2.2.2.2.2 Lift

Lift allows you to migrate on-premise data to cloud databases in a very efficient manner

More Information:
https://console.bluemix.net/catalog/services/lift-cli

2.2.2.2.3 Rocket Mainframe Data

Rocket Mainframe Data is leveraging similar functionality for batch style data integration like Lift but dedicated to IBM Mainframes

More Information:
https://console.bluemix.net/catalog/services/rocket-mainframe-data

2.3 Streaming analytics

2.3.1 Definition

The current state of the art is batch processing – since decades. But sometimes the value of a data product can be increased tremendously by adding real time analytics capabilities – since most of world's data loses value within seconds. Think of stock market data or the fact that a vehicle camera captures a pedestrian crossing a street. A streaming analytics system allows for real-time data processing. Think of it like running data against a continuous query instead of running a query against a finite data set.

2.3.2 Architectural Decision Guidelines

There exists a relatively limited set of technologies for real-time stream processing. The most important questions to be asked are:

- What throughput is required?
- What latency is accepted?
- Which data types must be supported?
- What type of algorithms run on the system? Only relational algebra or advanced modeling?
- What's the variance of the workload / what are the elasticity requirements?
- What type of fault tolerance / delivery guarantees are necessary?

2.3.2.1 Technology Guidelines

On the IBM cloud, there exist many service offerings for real-time data processing which will be explained in the following sections paired with guidelines when to use which one.

2.3.2.1.1 Apache Spark and Apache Spark Structured Streaming

Apache Spark is often the primary choice when it comes to cluster grade data processing and machine learning. If already used for batch processing – Apache Spark Structured Streaming should be the first thing to evaluate. This way, technology homogeneity can be achieved. Also, batch and streaming jobs can be run together (e.g. joining a stream of records against a reference table).

- What throughput is required?
  Apache Spark Structured Streaming supports the same throughput as in batch mode

- What latency is accepted?
  In Apache Spark v2.3 the Continuous Processing mode has been introduces, brining latency down to one millisecond

- Which data types must be supported?
  Apache Spark is string at structured and semi-structured data. Audio and Video data can't benefit from Apache Spark's accelerators "Tungsten" and "Catalyst"

- What type of algorithms run on the system? Only relational algebra or advanced modeling?
  Apache Spark Structured Streaming supports relational queries as well as machine learning – but machine learning is only supported on sliding and tumbling windows

- What's the variance of the workload / what are the elasticity requirements?
  Through their fault tolerant nature - Apache Spark clusters can be grown and shrinked dynamically

- What type of fault tolerance / delivery guarantees are necessary?
  Apache Spark Structured Streaming support exactly once delivery guarantees and depending on the type of data source complete crash fault tolerance.

2.3.2.1.2 IBM Streams

IBM Streams is the fastest streaming engine on the planet. Originally designed for low-latency, high throughput network monitoring applications IBM streams has its roots in cybersecurity.

- What throughput is required?
  IBM Streams is the fastest streaming engine on the market

- What latency is accepted?
  IBM Streams latency goes down to microseconds

- Which data types must be supported?
  Through IBM Streams binary transfer mode any type of data type can be supported

- What type of algorithms run on the system? Only relational algebra or advanced modeling?
  IBM Streams in its core supports all relation algebra. In addition, through toolkits various machine learning algorithms can be used. Toolkits are an open sytem and 3rdparty toolkits exists in the open source

- What's the variance of the workload / what are the elasticity requirements?
  Through its fault tolerant nature – IBM Streams clusters can be grown and shrinked dynamically

- What type of fault tolerance / delivery guarantees are necessary?
  IBM Streams supports exactly once delivery guarantees and complete crash fault tolerance

2.3.2.1.3 NodeRED

NodeRED is the most lightweight type of streaming engine. Implemented on top of Node.js in JavaScript it can even run on a 64 MB memory footprint (e.g. running it on a Raspberry PI)

- What throughput is required?
  NodeRED's throughput is bound to processing capabilities of a single CPU core though Node.js's event processing nature. For increased throughput, multiple instances of NodeRED have been used in parallel. Parallelization is not buil-in and needs to be provided by the application developer (e.g. round robin)

- What latency is accepted?
  Latency is also dependent on the CPU configuration and on the thoughphut sine high throughput congests the event queue and increases latency

- Which data types must be supported?
  NodeRED best supports JSON streams although any data type can be nested into JSON

- What type of algorithms run on the system? Only relational algebra or advanced modeling?
  NodeRED has one of the most extensive ecosystem of open source 3rd party modules. Although advance machine learning is not supported natively, there are plans by IBM to add those

- What's the variance of the workload / what are the elasticity requirements?
  Since parallelizing is a responsibility of the application developer, for independent computation a round-robin load balancing scheme supports linear scalability and full elasticity

- What type of fault tolerance / delivery guarantees are necessary?
  NodeRED has no built-in fault tolerance and no delivery guarantees

2.3.2.1.4 Apache Nifi

Apache Nifi is maintained by Hortonworks and part of the IBM Analytics Engine Service.

- What throughput is required?
  Nifi can handle hundreds of MB/s on a single node and can be configured to handle multiple GB/s in cluster mode

- What latency is accepted?
  Nifi's latency is in the second's range. Through message periodization the tradeoff between throughput and latency can be tweaked

- Which data types must be supported?
  Nifi best supports structured data streams although any data type can be nested into

- What type of algorithms run on the system? Only relational algebra or advanced modeling?
  Nifi supports relational algebra out of the box but custom processors can be built

- What's the variance of the workload / what are the elasticity requirements?
  Nifi can be easily scaled up without restarts, but scaling down requires stopping and starting the Nifi System

- What type of fault tolerance / delivery guarantees are necessary?
  Nifi supports end to end guaranteed exactly once delivery. Also fault tolerance can be configured, but automatic recovery is not possible. Another important feature is backpressure and pressure release which causes the upstream nodes to stop accepting new data and discarding unprocessed data if an age threshold is exceeded

2.3.2.1.5 Others

Among those technologies introduced exist a variety of others like Apache Kafka, Samza, Apache Flink, Apache Storm, Total.js Flow, Eclipse Kura, Flogo which might be worth to have a look at in case the ones above are not meeting all requirements

2.4 Data Integration

2.4.1 Definition

In the data integration stage data is cleansed, transformed and if possible, downstream features are added

2.4.2 Architectural Decision Guidelines

There exists an extremely huge set of technologies for batch data processing, which is the technology in use for data integration. The most important questions to be asked are:

- What throughput is required?
- Which data types must be supported?
- What source systems must be supported?
- What skills are required?

2.4.2.1 Technology Guidelines

On the IBM cloud, there exist many service offerings for data integration which will be explained in the following sections paired with guidelines when to use which one.

2.4.2.1.1 Apache Spark

Apache Spark is often the primary choice when it comes to cluster grade data processing and machine learning. Apache Spark is often a very flexible choice which also supports writing up integration processes in SQL. But a UI is missing.

- What throughput is required?
  Apache Spark scales linearly, so throughput is just a function of cluster size

- Which data types must be supported?
  Apache Spark works best with structured data but binary data is supported as well

- What source systems must be supported?
  Apache Spark can access a variety of SQL and NoSQL data based as well as file source out of the box. A common data source architecture allows adding capabilities. 3rd party project add functionality as well

- What skills are required?
  At least advanced SQL skills are required and some familiarity with either Java, Scala or python

2.4.2.1.2 IBM Data Stage on Cloud

IBM Data Stage is one of the most sophisticated ETL (Extract Transform Load) tools on the market. It's closed source and supports visual editing.

- What throughput is required?
  Data Stage can be used in cluster mode which supports scale-out

- Which data types must be supported?
  Data Stage has its roots in traditional Data Warehouse ETL and therefore concentrates on structured data

- What source systems must be supported?
  Again, Data Stage concentrates on relational database systems but files can also be read, even on Object Store. In addition, data sources can be added using plugins implemented in Java

- What skills are required?
  Since Data Stage is a visual editing environment, the learning curve is very low. No programming skills are required

2.4.2.1.3 Others

It's important to notice that data integration is mostly done using ETL tools or plain SQL or a combination of both. ETL tools are very mature technology and an abundance of technologies exist. On the other hand, if streaming analytics is part of the project it's worth to check if one of those technologies fits the requirements since reuse of such a system reduces technology heterogeneity with all its advantages.

2.5 Data Repository

2.5.1 Definition

This is the persistent storage for your data

2.5.2 Architectural Decision Guidelines

There exists an extremely huge set of technologies for persisting data. Most of them are relational databases. The second largest group are NoSQL databases and file system (including Cloud Object Store) form the last one. The most important questions to be asked are:

- How does is the impact of storage cost?
- Which data types must be supported?
- How good must point queries (on fixed or dynamic dimensions) be supported?
- How good must range queries (on fixed or dynamic dimensions) be supported?
- How good must full table scans be supported?
- What skills are required?
- What's the requirement for fault tolerance and backup?
- What are the constant and peak ingestion rates?
- What's the amount of storage needed?
- How does the growth pattern look like?
- What are the retention policies?

2.5.2.1 Technology Guidelines

On the IBM cloud, there exist many service offerings for SQL, NoSQL and file storage, which will be explained in the following sections paired with guidelines when to use which one.

2.5.2.1.1 Relational databases (RDBMS)

Dash DB is the DB2 BLU on the cloud offering from IBM featuring column store, advanced compression and execution on SIMD instructions sets (a.k.a. vectorized processing). But there exists also a variety of other options in the IBM Cloud like Informix, postgresql and MySQL

- How does is the impact of storage cost?
  Relational databases (RDBMS) have the highest requirements on storage quality. Therefore, cost if relational storage is always the highest

- Which data types must be supported?
  Relational databases are meant for structured data. Although there exist column data types for binary data which can be swapped out to cheaper storage, this is just an add-on and not core functionality of relational databases

- How good must point queries (on fixed or dynamic dimensions) be supported?
  RDBMS are king at point queries because an index can be created on each column

- How good must range queries (on fixed or dynamic dimensions) be supported?
  RDBMS are king at range queries because an index can be created on each column

- How good must full table scans be supported?
  RDBMS are trying to avoid full table scans in their SQL query optimizers. Therefore, performance is not optimized for full table scans (e.g. contaminating page caches)

- What skills are required?
  SQL skills are required for the application developer and if a cloud offering isn't chosen, data base administrator (DBA) skills are needed for the specific database

- What's the requirement for fault tolerance and backup?
  RDMBS support continuous backup and crash fault tolerance. For recovery, the system might need to go offline

- What are the constant and peak ingestion rates?
  Inserts using SQL are relatively slow, especially if the target table contains many indexes which must be rebalanced and updated as well. Some RDBMS support bulk inserts from files by bypassing the SQL engine but then the table usually needs to go offline for that period

- What's the amount of storage needed?
  RDMBS perform very well to around 1 TB of data. Going beyond that is complex and needs advanced cluster setups

- How does the growth pattern look like?
  RDBMS support volume management, so continuous growth usually isn't a problem, even during runtime. For shrinking the system might need to be taken offline

- What are the retention policies?
  RDBMS usually support automated retention mechanisms to delete old data automatically

2.5.2.1.2 NoSQL databases

The most prominent NoSQL databases like Apache CouchDB, MongoDB, Redis, RethinkDB, ScyllaDB (Cassandra) and InfluxCloud are supported.

- How does is the impact of storage cost?
  NoSQL databases are usually storage fault-tolerant by default. Therefore, quality requirements on storage is less which brings down storage cost

- Which data types must be supported?
  Although NoSQL databases are meant for structured data as well, they usually use JSON as storage format which can be enriched with binary data. Although, lot of binary data attached to JSON document can bring the performance down as well.

- How good must point queries (on fixed or dynamic dimensions) be supported?
  Some NoSQL databases support the creation of indexes which improve point query performance

- How good must range queries (on fixed or dynamic dimensions) be supported?
  Some NoSQL databases support the creation of indexes which improve range query performance

- How good must full table scans be supported?
  NoSQL databases are performing very well at full table scans. The performance is only limited by the I/O bandwidth to storage

- What skills are required?
  Usually, special query language skills are required for the application developer and if a cloud offering isn't chosen, data base administrator (DBA) skills are needed for the specific database

- What's the requirement for fault tolerance and backup?
  NoSQL databases support backups in different ways. But some aren't supporting online backup. NoSQL databases are usually crash fault tolerant, but for recovery, the system might need to go offline

- What are the constant and peak ingestion rates?
  Since usually, no indexes need to be updated and data doesn't need to be mapped to pages, ingestion rate are usually only bound to I/O performance of the storage system

- What's the amount of storage needed?
  RDMBS perform well to around 10 -100 TB of data. Cluster setups on NoSQL databases are much more straightforward than on RDBMS. Successful setups with >100 nodes and > 100.000 database reads/writes per second have been reported

- How does the growth pattern look like?
  Growth of NoSQL database is not a problem. Volumes can be added during runtime. For shrinking the system might need to be taken offline

- What are the retention policies?
  NoSQL databases don't support automated retention mechanisms to delete old data automatically, therefore this must be implemented manually resulting in range queries on the data corpus

2.5.2.1.3 Object Store

Cloud Object Store (OS) is to most disrupting storage technology in this decade, so let's have a look.

- How does is the impact of storage cost?
  Object Store is the cheapest option for storage

- Which data types must be supported?
  Since OS resembles a file system, any data type is supported

- How good must point queries (on fixed or dynamic dimensions) be supported?
  Since OS resembles a file system, external indices need to be created. It is possible to access specific storage locations through folder/file names and file offsets though

- How good must range queries (on fixed or dynamic dimensions) be supported?
  Since OS resembles a file system, external indices need to be created. It is possible to access specific storage locations through folder/file names and file offsets though. Therefore, range queries on a single defined column (e.g. data) can be achieved through hierarchical folder structures

- How good must full table scans be supported?
  Full table scans are just bound by the I/O bandwidth of the OS

- What skills are required?
  On a file level working with OS is much like working with any file system. Through Apache SparkSQL and IBM Cloud SQL Query data in OS can be accessed with SQL. Since OS is a cloud offering, no administrator skills are required. IBM OS is available for on-prem as well using an appliance box

- What's the requirement for fault tolerance and backup?
  Fault tolerance and backup is completely handled by the cloud provider. OS support intercontinental data center replication for high-availability out of the box.

- What are the constant and peak ingestion rates?
  Ingestion rates to OS is bound by the uplink speed to the OS system.

- What's the amount of storage needed?
  OS scale to the petabyte range

- How does the growth pattern look like?
  Growth and shrinking on OS is fully elastic

- What are the retention policies?
  Retention of data residing in OS must be done manually. Hierarchical File/Folder layout which is based on data/time helps here. Some OS support automatic movement of infrequently accessed files to colder storage (colder means less cost, but also less performance, or even higher cost of accesses to files)

2.6 Discovery and Exploration

2.6.1 Definition

This component allows for visualization and creation of metrics of data

Architectural Decision Guidelines
In various process models, data visualization and exploration is one of the first steps. Similar tasks are also applied in traditional data warehousing and business intelligence. So for choosing a technology, the following questions should be kept in mind:

- What type of visualizations are needed?
- Are interactive visualizations needed?
- Are coding skills available / required?
- What metrics can be calculated on the data?
- Do metrics and visualization need to be shared with business stakeholders?

2.6.1.1 Technology Guidelines
On the IBM cloud, there exist many service offerings data exploration. Some of those are open source, some aren't.

2.6.1.1.1 Jupyter, python, pyspark, scikit-learn, pandas, matplotlib, pixiedust
The components mentioned above are all open source and supported in the IBM Cloud. Some of them have overlapping features, some of them have complementary features. This will be made clear by answering the architectural questions

- What type of visualizations are needed?
  Matplotlib supports the widest range of possible visualizations including run chars, histograms, box-plots and scatter plots. Pixiedust as of V1.1.11 supports tables, bar charts, line charts, scatter plots, pie charts, histograms and maps

- Are interactive visualizations needed?
  Whereas matplotlib creates static plots, pixiedust supports interactive ones

- Are coding skills available / required?
  Whereas matplotlib needs coding skills, pixiedust does not. For computing metrics, some code is necessary

- What metrics can be calculated on the data?
  Using scikit-learn and pandas, all state-of-the-art metrics are supported

- Do metrics and visualization need to be shared with business stakeholders?
  Watson Studio supports sharing of jupyther notebooks, also using a fine grained user and access management system

2.6.1.1.2 SPSS Modeler
SPSS Modeler is available in the cloud and also as standalone product.

- What type of visualizations are needed?
  SPSS Modeler supports the following visualizations out of the box:

    - Bar
    - Pie
    - 3D Bar
    - 3D Pie
    - Line
    - Area
    - 3D Area
    - Path
    - Ribbon
    - Surface
    - Scatter
    - Bubble
    - Histogram
    - Box
    - Map

  More information can be found here
  https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/graphboard_creating_types.htm

- Are interactive visualizations needed?
  SPSS Modeler Visualizations are not interactive

- Are coding skills available / required?
  SPSS Modeler doesn't require any coding skills

- What metrics can be calculated on the data?
  All state-of-the-art metrics are supported using the Data Audit node

- Do metrics and visualization need to be shared with business stakeholders?
  Watson Studio supports sharing of SPSS Modeler Flows, also using a fine-grained user and access management system. But those might not be suitable to stakeholders

2.6.1.1.3 IBM Watson Analytics

IBM Watson Analytics is a complete SaaS (Software as a Service) offering for fully automated data exploration, visualization and modeling

- What type of visualizations are needed?
  IBM Watson Analytics supports the following out of the box:

    - Bar
    - Decision Tree
    - Spiral
    - Area
    - Bubble
    - Combination
    - Dial

    - Grid
    - Heatmap
    - Network
    - Pie
    - Table
    - Treemap
    - Word cloud

  More information can be found here
  https://www.ibm.com/watson-analytics

- Are interactive visualizations needed?
  IBM Watson Analytics not only supports interactive visualizations but also interactive creation and mortifications of visualizations

- Are coding skills available / required?
  IBM Watson Analytics a completely UI based solution

- What metrics can be calculated on the data?
  Although all state-of-the-art metrics are supported using the "Summary" functionality, this tool targets business users, so Data Scientists might get lost in abstractions

- Do metrics and visualization need to be shared with business stakeholders?
  IBM Watson Analytics is a collaborative platform with high level functionality perfectly suited to present to business stakeholders

2.7 Actionable Insights

2.7.1 Definition

This is where most of your work fits in. Here you create and evaluate your machine learning and deep learning models

2.7.2 Architectural Decision Guidelines

There exists an extremely huge set of technologies for creation and evaluation of machine learning and deep learning models. Although different technologies differ also in function and performance, those differences are usually neglectable. Questions which should be therefore asked are:

- What are the available skills regarding programming languages?
- What are the cost of skills regarding programming languages?
- What are the available skills regarding frameworks?
- What are the cost of skills regarding frameworks?
- Is model interchange required?
- Is parallel or GPU based training or scoring required?
- Do algorithms need to be tweaked or new algorithms to be developed?

2.7.2.1 Technology Guidelines

There exists an abundance of open and closed source technologies. Here, the most relevant are introduced. Although it holds for other sections as well, decisions made in this section are very prone to change due to the iterative nature of this process model. Therefore, changing or combining multiple technologies is no problem, although decisions let to those changes should be explained and documented.

2.7.2.1.1 IBM Watson Analytics

The SaaS offering already has been introduced. Watson Analytics has automated modeling capabilities which makes it a candidate here as well.

- What are the available skills regarding programming languages?
  As a complete UI based offering, Watson Analytics doesn't need any programming skills

- What are the cost of skills regarding programming languages?
  As a complete UI based offering, Watson Analytics doesn't need any programming skills

- What are the available skills regarding frameworks?
  Usually, there are no skills available within the team, but the learning curve is very steep

- What are the cost of skills regarding frameworks?
  As the learning curve is steep, costs of learning the tool is low

- Is model interchange required?
  Watson Analytics doesn't support import/export of models

- Is parallel or GPU based training or scoring required?
  Watson Analytics takes care of scaling automatically

- Do algorithms need to be tweaked or new algorithms to be developed?
  Watson Analytics doesn't allow for tweaking or introducing new algorithms which aren't supported by the tool

2.7.2.1.2 SPSS Modeler

SPSS Modeler already has been introduced. It really shines when it comes to machine learning.

- What are the available skills regarding programming languages?
  As a complete UI based offering, SPSS doesn't need programming skills, although it can be extended using R scripts

- What are the cost of skills regarding programming languages?
  As a complete UI based offering, SPSS doesn't need programming skills, although it can be extended using R scripts

- What are the available skills regarding frameworks?
  SPSS is one of the industry leader and skills are generally available

- What are the cost of skills regarding frameworks?
  Expert costs are usually lower in UI based tools than in programming frameworks

- Is model interchange required?
  SPSS Modeler supports PMML

- Is parallel or GPU based training or scoring required?
  SPSS Modeler supports scaling through IBM Analytics Server or IBM Watson Studio using Apache Spark

- Do algorithms need to be tweaked or new algorithms to be developed?
  SPSS Modeler algorithms can't be changed but using the R language algorithms can be added and custom developed

2.7.2.1.3 R/R-Studio

R and R-Studio have been the de-factor standard for open source based data science. It's supported in the IBM Cloud via IBM Watson Studio as well

- What are the available skills regarding programming languages?
  R programming skills are usually widely available since it is the standard programming language in many natural science based university curriculums. It can be acquired rapidly since it is just a procedural language with limited functional programming support.

- What are the cost of skills regarding programming languages?
  Costs of R programming are usually low

- What are the available skills regarding frameworks?
  R is not only a programming language but also requires knowledge of tooling (R-Studio) and especially knowledge of the very abundant R library (CRAN) with 6000+ packages

- What are the cost of skills regarding frameworks?
  Expert costs are correlated with knowledge of the CRAN library and years of experience and in the range of usual programmer costs

- Is model interchange required?
  Some R libraries support exchange of models but it is not standardized

- Is parallel or GPU based training or scoring required?
  Some R libraries support scaling and GPU acceleration but it is not standardized

- Do algorithms need to be tweaked or new algorithms to be developed?
  R needs algorithms to be implemented in C/C++ to run fast. So, tweaking and custom development usually involved C/C++ coding

2.7.2.1.4 Python, pandas and scikit-learn

Although R and R-Studio have been the de-factor standard for open source based data science for a long period of time, Python, pandas and scikit-learn are the runner-up. Python is a much cleaner programming language than R and easier to learn therefore. Pandas is the python equivalent to R dataframes supporting relational access to data. Finally, scikit-learn nicely groups all necessary machine learning algorithms together. It's supported in the IBM Cloud via IBM Watson Studio as well.

- What are the available skills regarding programming languages?
  Python skills are very widely available since python is a clean and easy to learn programming language

- What are the cost of skills regarding programming languages?
  Because of python's properties mentioned above, cost of python programming skills are very low

- What are the available skills regarding frameworks?
  Pandas and scikit-learn are very clean and easy to learn frameworks, therefore skills are widely available

- What are the cost of skills regarding frameworks?
  Because of the properties mentioned above, cost of skills are very low

- Is model interchange required?
  All scikit-learn models can be (de)serialized. PMML is supported via 3rd party libraries

- Is parallel or GPU based training or scoring required?
  Neither GPU nor scale-out is supported, although scale-up capabilities can be added individually to make use of multiple cores

- Do algorithms need to be tweaked or new algorithms to be developed?
  Scikit-learn algorithms are very cleanly implemented. They all stick to the pipelines API making reuse and interchange easy. Linear algebra is handled throughout with the numpy library. So, tweaking and adding algorithms is straightforward

2.7.2.1.5 Python, Apache Spark and SparkML

Although python, pandas and scikit-learn are more widely adopted, the Apache Spark ecosystem is catching up. Especially because of its scaling capabilities. It's supported in the IBM Cloud via IBM Watson Studio as well.

- What are the available skills regarding programming languages?
  Apache Spark supports python, Java, Scala and R as programming languages

- What are the cost of skills regarding programming languages?
  The costs depend on what programming language is used with Python being usually the cheapest

- What are the available skills regarding frameworks?
  Apache Spark skills are on high demand and usually not available

- What are the cost of skills regarding frameworks?
  Apache Spark skills are on high demand and usually expensive

- Is model interchange required?
  All SparkML models can be (de)serialized. PMML is supported via 3rd party libraries

- Is parallel or GPU based training or scoring required?
  All Apache Spark jobs are inherently parallel. But GPU's are only supported through 3rd party libraries

- Do algorithms need to be tweaked or new algorithms to be developed?
  As in Scikit-learn, algorithms are very cleanly implemented. They all stick to the pipelines API making reuse and interchange easy. Linear algebra is handled throughout with built-in Apache Spark libraries. So, tweaking and adding algorithms is straightforward

2.7.2.1.6 Apache SystemML

When it comes to relational data processing, SQL is king. Mainly because an optimizer takes care on optimal query executions. Think of SystemML as an optimizer for linear algebra capable of creating optimal execution plans for jobs running on data parallel frameworks like Apache Spark

- What are the available skills regarding programming languages?
  SystemML has two domain specific languages (DSL) with R and python syntax

- What are the cost of skills regarding programming languages?
  Although the DSLs are like R and python, there is a learning curve involved

- What are the available skills regarding frameworks?
  SystemML skills are very rare

- What are the cost of skills regarding frameworks?
  Due to the high learning curve and skill scarcity costs might get high

- Is model interchange required?
  SystemML models can be (de)serialized. PMML is not supported. SystemML can import and run Caffe2 and Keras models

- Is parallel or GPU based training or scoring required?
  All Apache Spark jobs are inherently parallel. SystemML makes use of this property. In addition, GPU's are only supported as well.

- Do algorithms need to be tweaked or new algorithms to be developed?
  Although SystemML comes with a large set of pre-implemented algorithms for machine learning and deep learning, it's strengths are in tweaking existing algorithms or implementing new ones because the DSL allows for concentrating on the mathematical implementation of the algorithm, the rest is handled by the framework. This makes it an ideal choice for these kind of tasks

2.7.2.1.7 Keras and TensorFlow

TensorFlow is one of the most widely used DeepLearning frameworks. In its core, it is a linear algebra library supporting automatic differentiation. TensorFlow's python driven syntax is relatively complex. Therefore, Keras provides an abstraction layer on top of TensorFlow. Both framework are seamlessly supported in the IBM Cloud through Watson Studio and Watson Machine Learning

- What are the available skills regarding programming languages?
  Python is the core programming language for Keras and TensorFlow

- What are the cost of skills regarding programming languages?
  Python programmers are relatively cheap

- What are the available skills regarding frameworks?
  Keras and TensorFlow skills are very rare

- What are the cost of skills regarding frameworks?
  Due to the high learning curve and skill scarcity costs might get high

- Is model interchange required?
  Keras and TensorFlow have their own model exchange formats. There exist converters from and to ONNX

- Is parallel or GPU based training or scoring required?
  Running TensorFlow on top of ApacheSpark is supported through TensorFrames and TensorSpark. Keras models can be run on ApacheSpark using DeepLearning4J and SystemML. Both latter frameworks also support GPUs. TensorFlow (and therefore Keras) support GPU natively as well.

- Do algorithms need to be tweaked or new algorithms to be developed?
  TensorFlow is a linear algebra execution engine, therefore optimally suited for

tweaking and creating new algorithms. Keras is a very flexible DeepLearning library supporting all types of neural network layouts.

2.8 Applications / Data Products

2.8.1 Definition

Models are fine but their value rises when they can be consumed by the ordinary business user – therefore one needs to create a data product. Data Products doesn't necessarily need to stay on the cloud – they can be pushed to mobile or enterprise applications

2.8.2 Architectural Decision Guidelines

In contrast to machine learning and deep learning frameworks, the space of frameworks to create data product is tiny. This maybe reflects what the current state-of-the-art in data science concentrates on. Depending on the requirements, Data Products are relatively visualization centric after a lot of use input data has been gathered. They also might involve asynchronous workflows as batch data integration and model training/scoring is performed within the workflow. Questions which should be therefore asked are:

- What skills are present for developing a data product?
- What skill are necessary for developing a data product?
- Is instant feedback required or is batch processing accepted?
- What's the degree of customization needed?
- What's the target audience? Is cloud scale deployment for a public use base required?

2.8.2.1 Technology Guidelines

Currently, only a limited set of frameworks and technologies is available in different categories. Below, the most prominent examples are given.

2.8.2.1.1 R-Shiny

R-Shiny is the most famous framework for building data products. Closely tied to the R language it enables data scientists to rapidly create a UI on top of existing R-scripts.

- What skills are present for developing a data product?
  R-Shiny requires R development skills. So it best fits into a R development ecosystem

- What skill are necessary for developing a data product?
  Although based on R, R-Shiny needs additional skills. For experienced R developers, the learning curve is steep and additional knowledge to acquire is minimal

- Is instant feedback required or is batch processing accepted?
  The messaging model of R-Shiny supports instant UI feedback when server side

data structures are updated. So response time is independent of the framework and therefore should be considered and resolved programmatically on the server side

- What's the degree of customization needed?
  Although R-Shiny is an extensible framework, extending it requires a deep understanding of the framework and R-Technology. Out of the box, there is a (huge) set of UI widgets end elements supported allowing for very customizable applications. If requirements are going beyond those capabilities, costly extension is required.

- What's the target audience? Is cloud scale deployment for a public use base required?
  R-Shiny application look very professional, although quite distinguishable. Therefore, the target audience must accept the UI design limitations. R-Shiny is best dynamically scaled horizontally in a container environment like Kubernetes. Ideally, every user session runs in its own container since R and R-Shiny are very sensitive to main memory shortages

2.8.2.1.2 Node-RED

Although Node-RED is a No-Code/Low-Code data flow/data integration environment, due to its modular nature it supports various extensions including the dash boarding extension. This extension allows for fast creation of user interfaces including advanced visualizations which are updated in real-time.

- What skills are present for developing a data product?
  Due to the completely graphical user interface based software development approach, only basic skills are required to build data products with Node-RED

- What skill are necessary for developing a data product?
  Any resource familiar with flow based programming as used in many state-of-the-art ETL and data mining tools will have a fast start with Node-RED. Basic JavaScript knowledge is required for creating advanced flows and for extending the framework

- Is instant feedback required or is batch processing accepted?
  The UI instantly reflects updates of the data model. Therefore, all considerations regarding feedback delay should be considered when developing the data integration flow or potentially involved calls to synchronous or asynchronous 3rd party services

- What's the degree of customization needed?
  Node-RED is a Node.js/JavaScript based framework. Custom UI widget require advanced Node.js development skills

- What's the target audience? Is cloud scale deployment for a public use base required?

Node-RED dashboard can be deployed for a pubic user base as long as the limitation regarding UI customization are acceptable. Sine Node.js runs on a single threaded event loop, scaling must be done horizontally, preferably using a containerized environment. Note: The Internet of Things Starter kit in the IBM Cloud supports horizontal scaling out of the box

2.8.2.1.3 D3

When it comes to custom application development, D3 is the most prominent and most widely used visualization widget framework which a very large open source ecosystem contribution reams of widgets for every desirable use case

- What skills are present for developing a data product?
  D3 fit's best into an existing, preferably JavaScript based developer ecosystem, although JavaScript is only required at client side. Therefore, at server side any REST based endpoints in any programming language are supported. Once example is REST endpoints accessed by a D3 UI provided by Apache Livy which encapsulates Apache Spark jobs

- What skill are necessary for developing a data product?
  D3 requires sophisticated D3 and at least client side JavaScript skills. Skills in a JavaScript AJAX framework like AngularJS are highly recommended. On the server side, capabilities of providing REST endpoints to the D3 applications are required

- Is instant feedback required or is batch processing accepted?
  The UI instantly reflects updates of the data model. Therefore, all considerations regarding feedback delay should be considered when developing the data integration flow or potentially involved calls to synchronous or asynchronous 3rd party services

- What's the degree of customization needed?
  D3 applications usually are implemented from scratch anyway. Therefore, this solution provides the most flexibility to the end user

- What's the target audience? Is cloud scale deployment for a public use base required?
  As a cloud native application, a D3 based data product can provide all capabilities for horizontal and vertical scaling and full adoption to user requirements

2.9 Security, Information Governance and Systems Management

2.9.1 Definition

This important step is forgotten easily – it's important to control who has access to which information for many compliance regulations. In addition, modern data science architectures involve many components which require operational aspects as well.

2.9.2 Architectural Decision Guidelines

Data privacy is one major challenge in many data science projects. Questions which should be therefore asked are:

- What granularity is required for managing user access to data assets?
- Are existing user registries required to be integrated?
- Who is taking care about operational aspects?
- What are the requirements for data retention?
- What level of security against attacks from hackers is required?

2.9.2.1 Technology Guidelines

Again, there exists an abundance of software and ways to solve the requirements involved in this topic and representative examples are chosen as exemplars.

2.9.2.1.1 Internet Services

Deploying a productive client facing web applications brings along serious risks. IBM Cloud Internet Services provides Global Points of Presence (PoPs). It includes Domain Name Service (DNS), Global Load Balancer (GLB), Distributed Denial of Service (DDoS) protection, Web Application Firewall (WAF), Transport Layer Security (TLS), and Caching

- What level of security against attacks from hackers is required?
  Internet Services is using services from CloudFlare, the world leader in this space

2.9.2.1.2 IBM App ID

Identity Management allows for cloud based user and identity management for web and mobile applications, APIs and backend systems. “Cloud users” can sign-up and sign-in with App ID's scalable user registry or social log-in with Google or Facebook. “Enterprise users” can be integrated using SAML 2.0 federation

- What granularity is required for managing user access to data assets?
  IBM App ID supports user management but no group/roles. Therefore, fine-grained access must be managed within the application

- Are existing user registries required to be integrated?
  IBM App ID supports registry integration using SAML 2.0 federation

2.9.2.1.3 Object Store

Object Store is the de-facto standard when it comes to modern, cost-effective cloud storage

- What granularity is required for managing user access to data assets?
  IBM Identity and Access Management (IAM) integration allows for granular access control at the bucket level using role-based policies

- What are the requirements for data retention?

Object Store supports different storage classes for frequently accessed data, occasionally accessed data and long-term data retention with Standard, Vault, and Cold Vault. The “Flex class” allows for dynamic data access and automates this process. Physical deletion of data still must be triggered externally

- Who is taking care about operational aspects?
  Regional and Cross Region resiliency options allow for increased data reliability

- What level of security against attacks from hackers is required?
  All data is encrypted at-rest and in-flight by default. Keys are automatically managed by default, but can optionally be self-managed or managed using IBM Key Protect

2.9.2.1.4 IBM Cloud PaaS/SaaS

IBM Cloud PaaS/SaaS eliminates operational aspects from data science project since all components involved are managed by IBM

- Who is taking care about operational aspects?
  In PaaS/SaaS clouds operational aspects are being taken care of by the cloud provider