The Lightweight IBM Cloud Garage Method for Data Science

Architectural Decisions Document Template

# Architectural Components Overview



IBM Data and Analytics Reference Architecture. Source: IBM Corporation

## Data Source a) Datastream **E1a** contains measured (Link to Datastream D) values of gas phase pollutants (e.g. Ozone, NO2, SO2, CO) measured in 2017 over Germany hourly: ([https://datahub.uba.de/server/rest/directories/arcgisforinspire/INSPIRE/aqd\_MapServer/Daten/AQD\_DE\_E1a\_2017.zip](https://datahub.uba.de/server/rest/directories/arcgisforinspire/INSPIRE/aqd_MapServer/Daten/AQD_DE_E1a_2017.zip" \t "_blank)). For model extension purposes additional data for years 2012-2018 (at least) can be found at the same place.

## b) Prevalence of Heart failures in the 2017: [https://www.versorgungsatlas.de/fileadmin/excel/data\_id\_97\_kreis11\_2\_j\_1483228800.xlsx](https://www.versorgungsatlas.de/fileadmin/excel/data_id_97_kreis11_2_j_1483228800.xlsx" \t "_blank) For model extension purposes additional data for years 2008-2017 and for other diseases can be found at the same place.

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.  
 The datasets are choosen in xml format for Datastream **E1a, and xlsx format for Heart Failure rates**

### Justification

Please justify your technology choices here.

These formats (xml and xlsx) are available and could be easily parsed by means of pandas.

## Enterprise Data

Independent Pre-processed datasets, including pollutant values pro hour pro sensor;   
 heart failure rates pro county; sensor ID/county table.

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Pandas data frames  
 Parquet storage

### Justification

Please justify your technology choices here.

Easy to pre-process and perform exploratory analysis   
 compact file size

## Streaming analytics

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Not needed

### Justification

Please justify your technology choices here.

New data becomes available in one chunk pro year.

## Data Integration

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Spark RDS object with SQL connector

### Justification

Please justify your technology choices here.

Fast access to big data set, using IBM Watson machine learning infrastructure.

## Data Repository

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

Object store

### Justification

Please justify your technology choices here.  
  
Good scalability with increasing the model size

## Discovery and Exploration

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

### Justification

Please justify your technology choices here.

## Actionable Insights

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

### Justification

Please justify your technology choices here.

## Applications / Data Products

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

### Justification

Please justify your technology choices here.

## Security, Information Governance and Systems Management

### Technology Choice

Please describe what technology you have defined here. Please justify below, why. In case this component is not needed justify below.

### Justification

Please justify your technology choices here.