Dimensions and Metrics

Nitin Verma
nitin@rightster.com

Introduction

Dimensions define points in euclidean space and metrics are the scalar quantities measured on the individual points in N-dimensional space, E_N . The same problem can formalized in a more generalized vector space. For now we can consider these as N-dimensional points and measurements on the those points.

Point in N-dimensional space

$\vec{d}=(d_1,d_2,\ldots,d_N)$ where $d_1,d_2,\ldots,d_N$ are one of the possible values in N dimensions respectively.

Above is a point represented as a vector from the origin.

Metric (measure) at $\vec{d}$

$m^{\vec{d}}$

This is a scalar quantity measured on that point.

Problem space

In problems we solve dimensions are discrete and finite at any given time. Those can be numbers, enumerations, time etc. So these always have a granularity defined and have finite possible values at any given time. The possible values may change across time. Time as dimension arguably is continues and has infinite possible values. Besides that fact we are presented with finite set of values with defined granularity (sec, days, months etc.) at any given time.

Simple model

Considering above facts we can safely model the problem statement as follows.

$$d_{n_k}\in D_n, D_n=\{d_{n_1},d_{n_2},\ldots,d_{n_K}\}:1\le k\le K;1\le n\le N:\{k,K,n,N\}\in\mathbb{Z}$$

d_{n_k} is the k^{th} possible value in n^{th} dimension.
D_n is the set of all the possible K values in n^{th} dimension.

K is also referred as **Cardinality** of the dimension, count of all the possible values that a dimension can have.

So the overall cardinality of a N-dimensional space, κ_N , is a product of cardinality in all N dimensions.

$$\kappa_N=\prod_{n=1}^{N}K_n$$

where K_n is cardinality in n^{th} dimension.

Direct Metric

Metric that is measure by observation. For example word count where word would

be a dimension.

Derived Metric

Metric that is calculated based on other metric, direct or derived.

$$m_1^{\vec{d}} = f(\{m^{\vec{d}} : m^{\vec{d}} \in \Gamma^{\vec{d}}; m_1^{\vec{d}} \notin \Gamma^{\vec{d}}; \Gamma \subset M^{\vec{d}}\})$$

Where,
$M^{\vec{d}}$ is the set of all metrics we are interested on a given point $\vec{d}$
$\Gamma^{\vec{d}}$ is a subset of $M^{\vec{d}}$ that essential does not include $m^{\vec{d}}$

This means derived metric are function of one or more metrics.

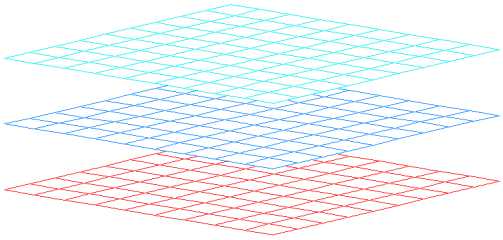
Fact tables

$M^{\vec{d}}$ is a single record in your fact table, thus fact tables can be defined as set of all measurements, $M^{\vec{d}}$ on observable points in an N-dimensional space.

Fact, $\tau_N = \{M^{\vec{d}} : \vec{d} \in E_N\}$

Where,
E_N is the set of observable points in the N-dimensional euclidean space.

Dimensional Aggregates



Assume we have three dimensions and third dimension has a cardinality of 3. In this case to aggregate to two dimensions, there needs to be an operation that converts these three planes to one.

In general case we'll operate on hyperplanes, I have taken liberty of three dimensions just for ease of presentation.

$$\tau_M = \Psi(\tau_N) : M < N; \{M, N\} \in \mathbb{Z}$$

Ψ is a function that operates on an higher order fact to produce a lower order fact.

Ψ can be further defined as a set of functions capable of producing individual metric in lower order.

Considering that,

$$\Psi(\tau_N) := \{\psi_{m_1}(\tau_N), \psi_{m_2}(\tau_N), \ldots, \psi_{m_A}(\tau_N)\}$$

every function produces an individual metric

$$m_\alpha^{\vec{d}_M} = \psi_{m_\alpha}(\tau_N) : 1 \le \alpha \le A$$

A is the number of direct and derived metric
$m_\alpha^{\vec{d}_M}$ is one of the metric calculated for a point $\vec{d}_M$ in M-dimensions

These are the functions we are most interested in. We should take greater care in understanding these functions in a particular problem. As normally ETLs tend to do multiple aggregations we should look at these closely.

Pose following question for every instance of such occurrence

If aggregation is performed to get one lower order dimension that is N to N-1

$$m_{\alpha}^{\vec{d_{N-1}}} = \psi_{m_{\alpha}}(\tau_N)$$

Would it yield the same results if aggregates are done progressive or direct.

$$m_{\alpha}^{\vec{d_{N-2}}} = \psi_{m_{\alpha}}(\tau_{N-1})$$
$$\mu_{\alpha}^{\vec{d_{N-2}}} = \psi_{m_{\alpha}}(\tau_N)$$

That is to analyze the above equations saying is $m_{\alpha}^{\vec{d_{N-2}}} = \mu_{\alpha}^{\vec{d_{N-2}}}$

Summation as an aggregate function passes this test, but average does not. Thus we have to analyze the same.

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$$

and we can prove

$$\frac{1}{n}\sum_{i=1}^{n} x_i \neq \frac{1}{n_1}\sum_{i=1}^{n_1} x_i + \frac{1}{n_2}\sum_{i=1}^{n_2} x_i : n = n_1 + n_2$$

Solution would be to maintain three metrics, two direct and one derived. As summation and count are both additive, we can safely progressively aggregate these and derive average.

For instance, if you had to deal with standard deviation, it would not be possible applying the standard form.

$$\sigma^2 := \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

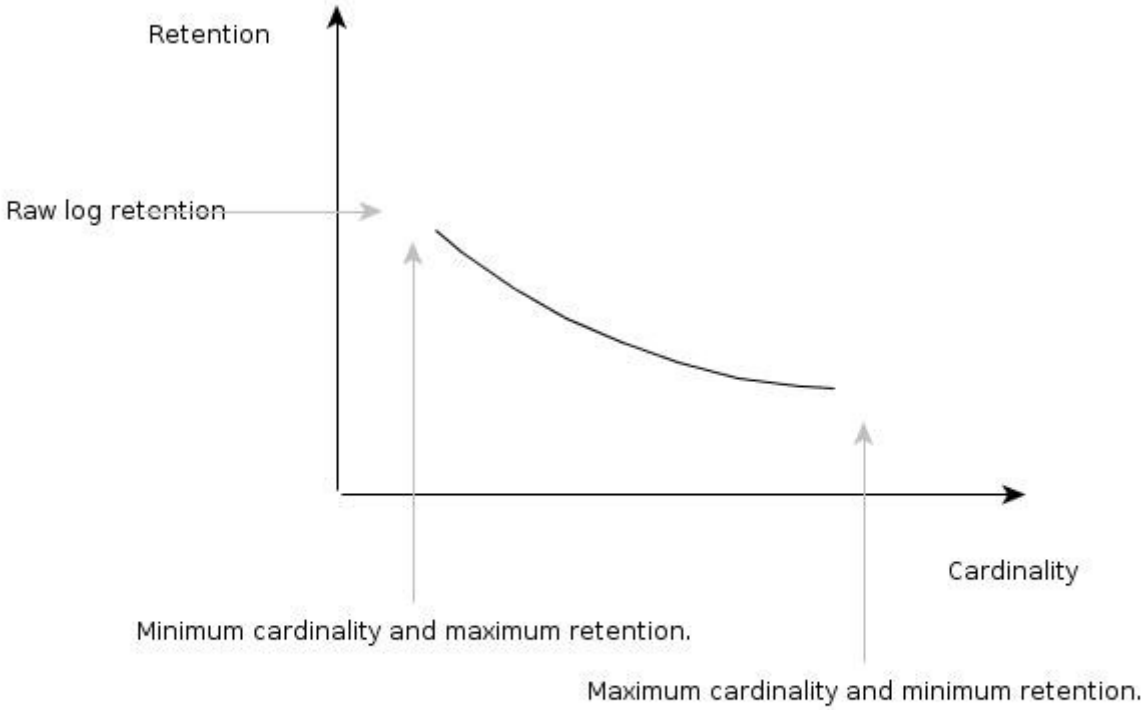
Computational form using König-Huygens theorem would suit in this case.

$$\sigma^2 := \overline{x^2} - \bar{x}^2$$

Give the above equation we need 4 metrics, two direct and two derived.

In situations where these kind of reductions are not possible, the aggregation is always performed for those metrics from highest order facts.

Data Retention



Retention x Cardinality = Constant