Dynamic Data Manager

Future algorithm proposal

Related Work

- Lots of similar work in the last years
- But most focus on the underlying services
- Want to focus on a balanced system
- When and where to replicate/delete datasets

Goals

$$Q_t = \frac{\sum\limits_{d \in \mathcal{D}} \frac{\sum\limits_{j_d \in \mathcal{J}_d} qt}{|\mathcal{J}_d|}}{|\mathcal{D}|} \tag{8}$$

- Minimize number of unused datasets (Victor?)

- Minimize median job waiting time (hard to measure, Equation 8)

- Balanced system in terms of accesses/GB (Equation 2)

Balance equations

$$\delta = \int_0^t \frac{n\_accesses}{size\_GB * n\_replicas} \, \mathrm{d}t \tag{1}$$

$$\sigma = \sqrt{\frac{\sum\limits_{d \in \mathcal{D}} (\delta_d - \overline{\delta})^2}{|\mathcal{D}| - 1}} \tag{2}$$

$$\delta_f = \delta - \overline{\delta} \tag{3}$$

Equation 1 measures the ratio of number of accesses per GB for one dataset during a time period.

Equation 2 measures the system-wide state of balance

Equation 3 gives a measurement of the balance for one dataset

“Data Dealer”

- Create a vector where each cell is the delta_f for one dataset
- A negative value means there are too many replicas
- A positive value means there isn’t enough replicas
- Run simulations to find out for which values it is worth making new replicas/delete

“Data Dealer” cont’d

- Avoid spikes to affect the algorithm
- Use delta_f from the n last days
- Use an exponential weight distribution

Site Selection

- Base decision on available CPU and storage
- CPU max = average top 3 days in last 3 weeks
- Some work have already been done here