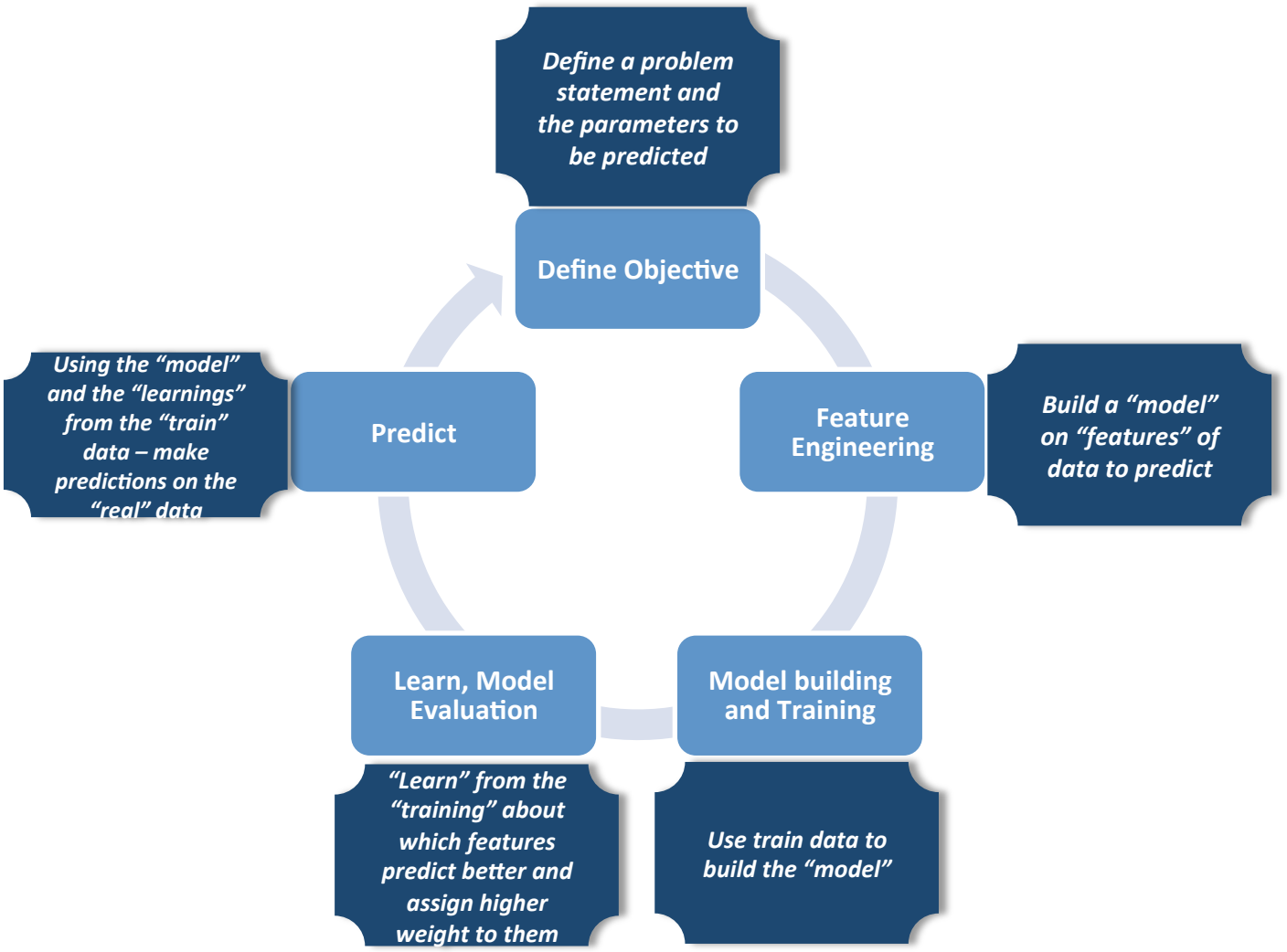
Predictive Analytics - 101



Predictive Learning - 101

Predict whether a logo is of American Express or not?

Define Objective

Feature Engineering

Key Features
-- color = blue;
-- It has some letters in white characters
--shape = square

Model building and Training

Build a Decision Tree with the above features on train data

Learn, Model Evaluation

New Learning :
-- shape is sometimes rectangle
-- color could also be a darker shade of blue

Predict

Is the logo on the left Amex?

Real time Spend intent prediction…

For each transaction.
- **Collate historic features**
- **Compute Dynamic Features**
- **Combine Static and Dynamic Features related to CM, SE and geo-location**
- **For each category execute the GBDT model In-parallel and compute the predictions**
- **If the predicted probability is greater than the threshold for that category, then, suggest the category as a possible category that the CM might visit next.**

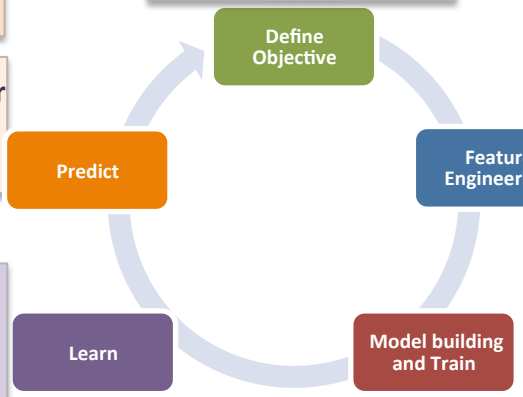
Peak traffic of 30 TPS equating to 30K lookups per second and 5K complex calculations per second; 900 GBDT executions in-parallel at peak

Through training rank order the features that predict better and assign each of these features some weights such that better predicting feats. Have more weights.
Eg. Phase_of_day might have lesser weights as compared to the feature indicating the Periodicity of the CM visiting a particular SE.

Calculate the prediction thresholds for each Category.
Eg. If prediction probability for Restaurant Category is greater than 0.1 then consider it true

~1000 feature collation and lookup; ~100 dynamic feature computation and running 30+ models

Predict if a CM will transact at a particular category in the next two hours



Data points for Modeling. (i) Transaction data (ii) SE Characteristics

Static Features
- Does the CM visit SE periodically?
- How much time does the CM spend in a category (Lifestyle, Entertainment, etc.)?
- CM's home location
- CM's FICO score
- CM's share of wallet

Dynamic Features
- Time of the day, day of the week, phase of day
- Trends in the neighboring merchants
- Distance of the CM from the home location
- Number of times the CM has transacted in a *session*

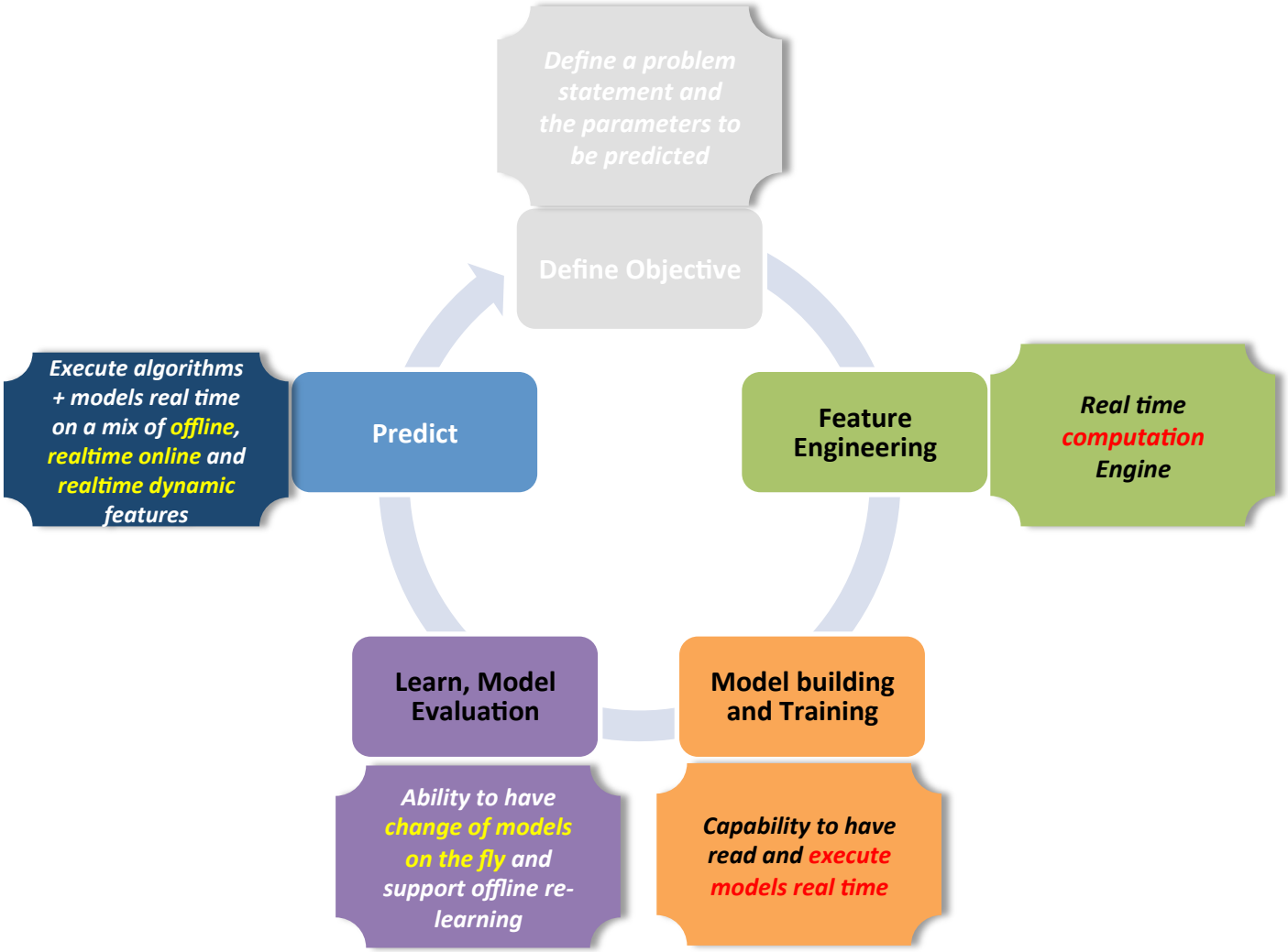
~60 GB of data size; ~1000 features,~100 dynamic 3MM CMs, 150K SEs. Facilitating sub-second search for all CMs and SEs

Build several decision on different subsets of the Features.
Eg. if (phase_of_day="morning") and (CM has not transacted at "restaurant" in this session) and (CM periodically visits "restaurants" before Noon); then – there is a 95% chance that the CM Will transact in the "RESTAURANT" category.

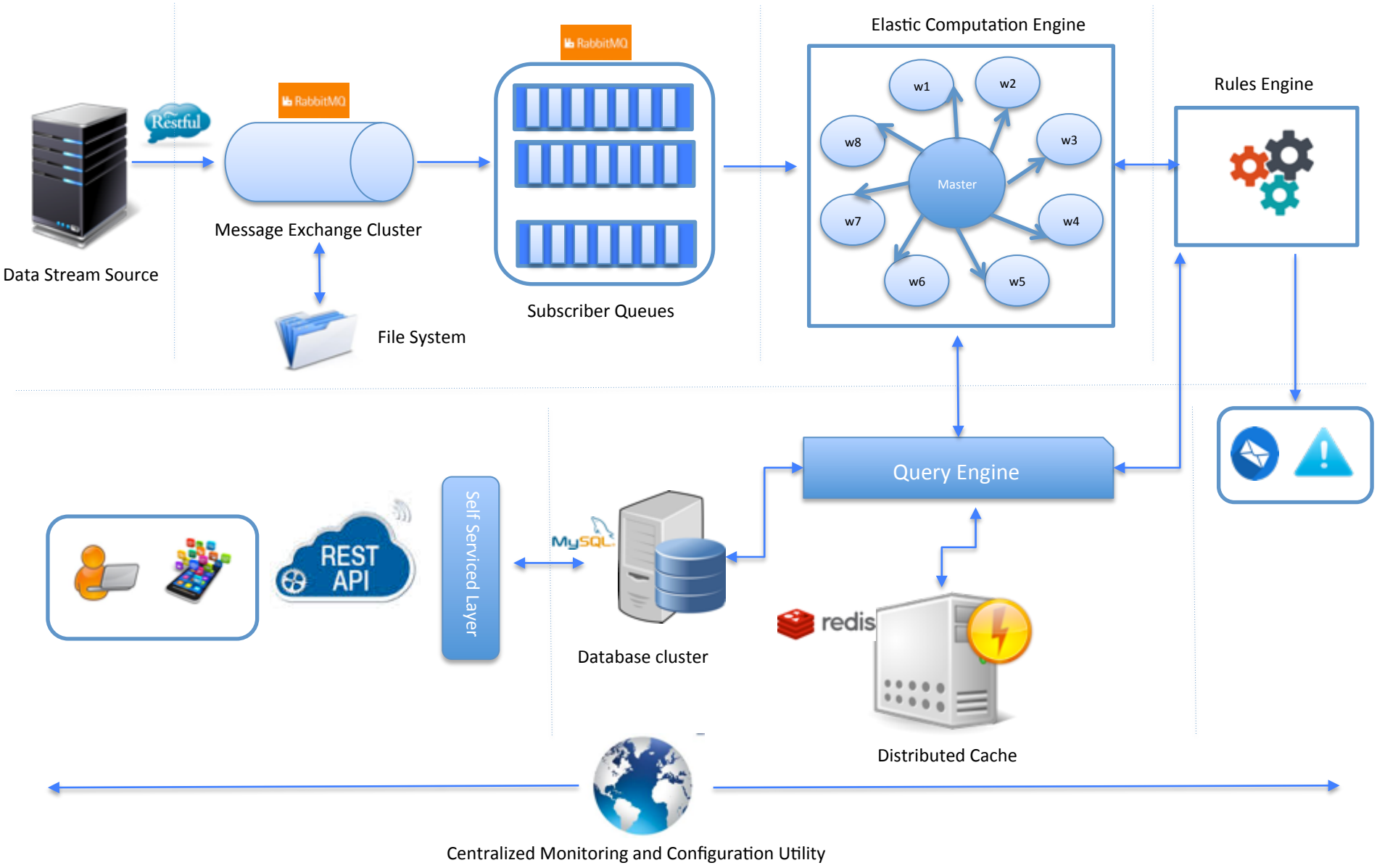
Create a Gradient Boosting Decision Trees(GBDT) model of the training set for each category of merchants

30 GBDT models with 200 Decision Trees each leading 30 parallel executions of GBDT ~900ms execution time

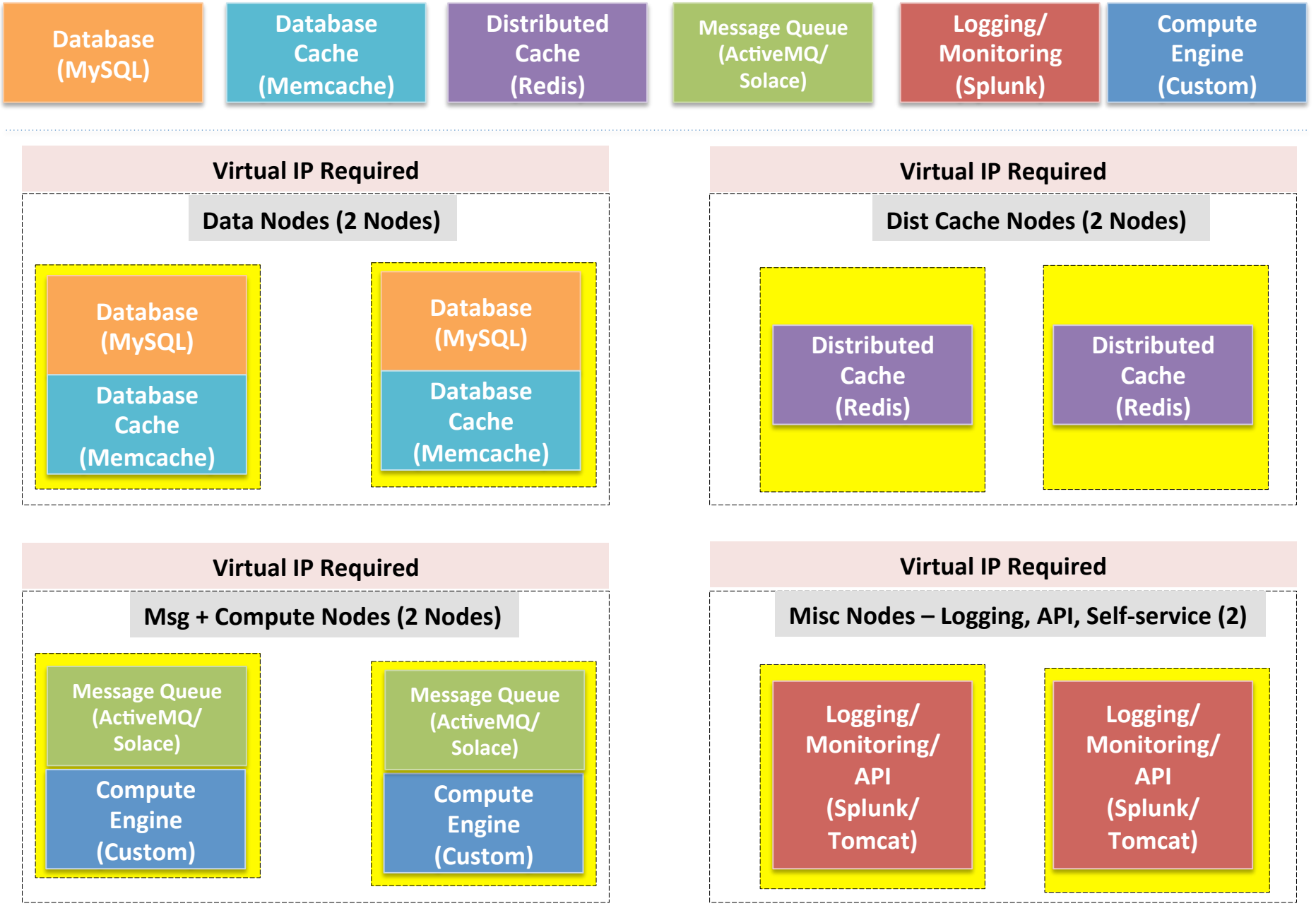
Real time Prediction at 30k feet…



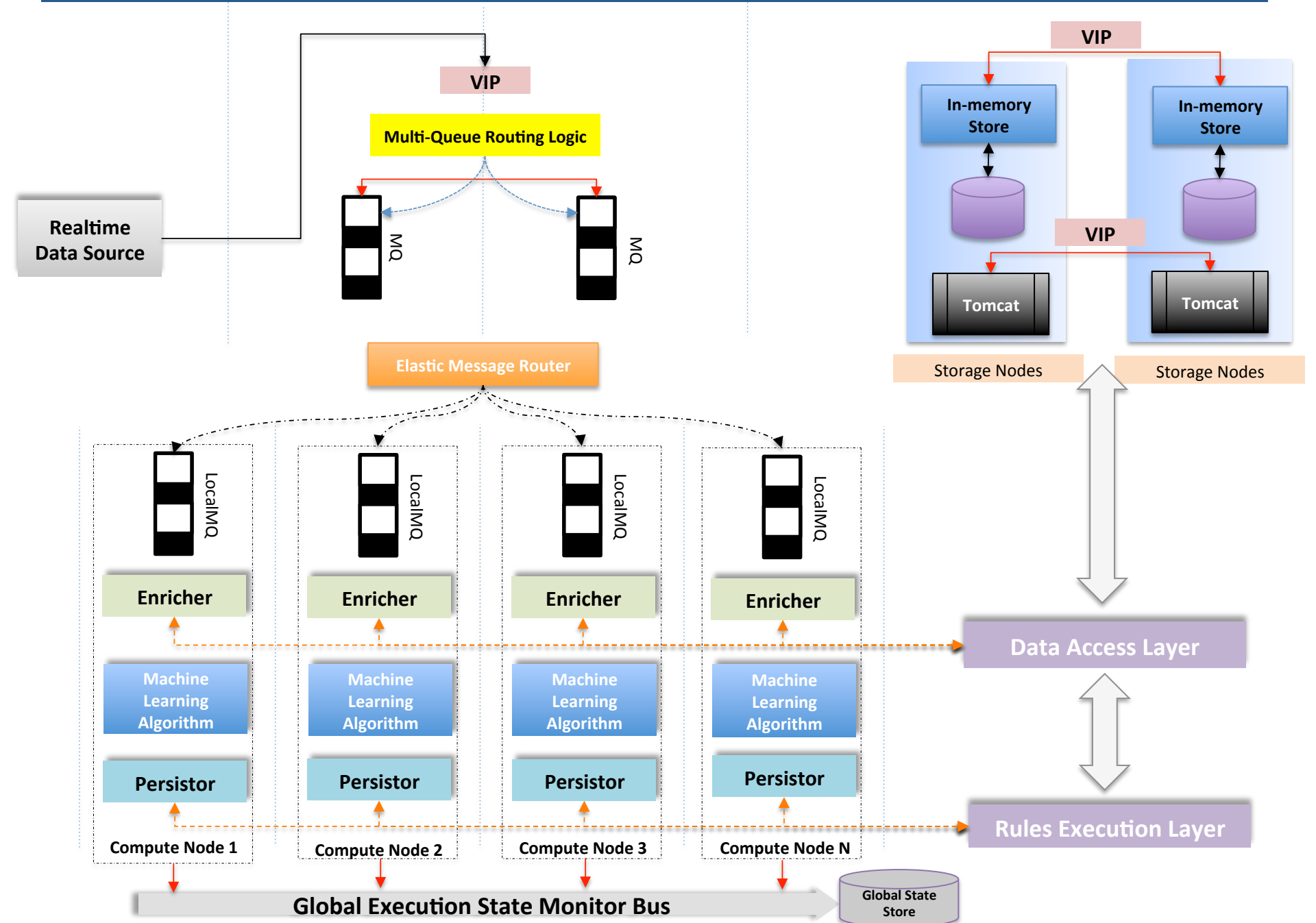
Big Data – Real Time Analytics Platform (CAPSTONE) : End To End Logical Architecture



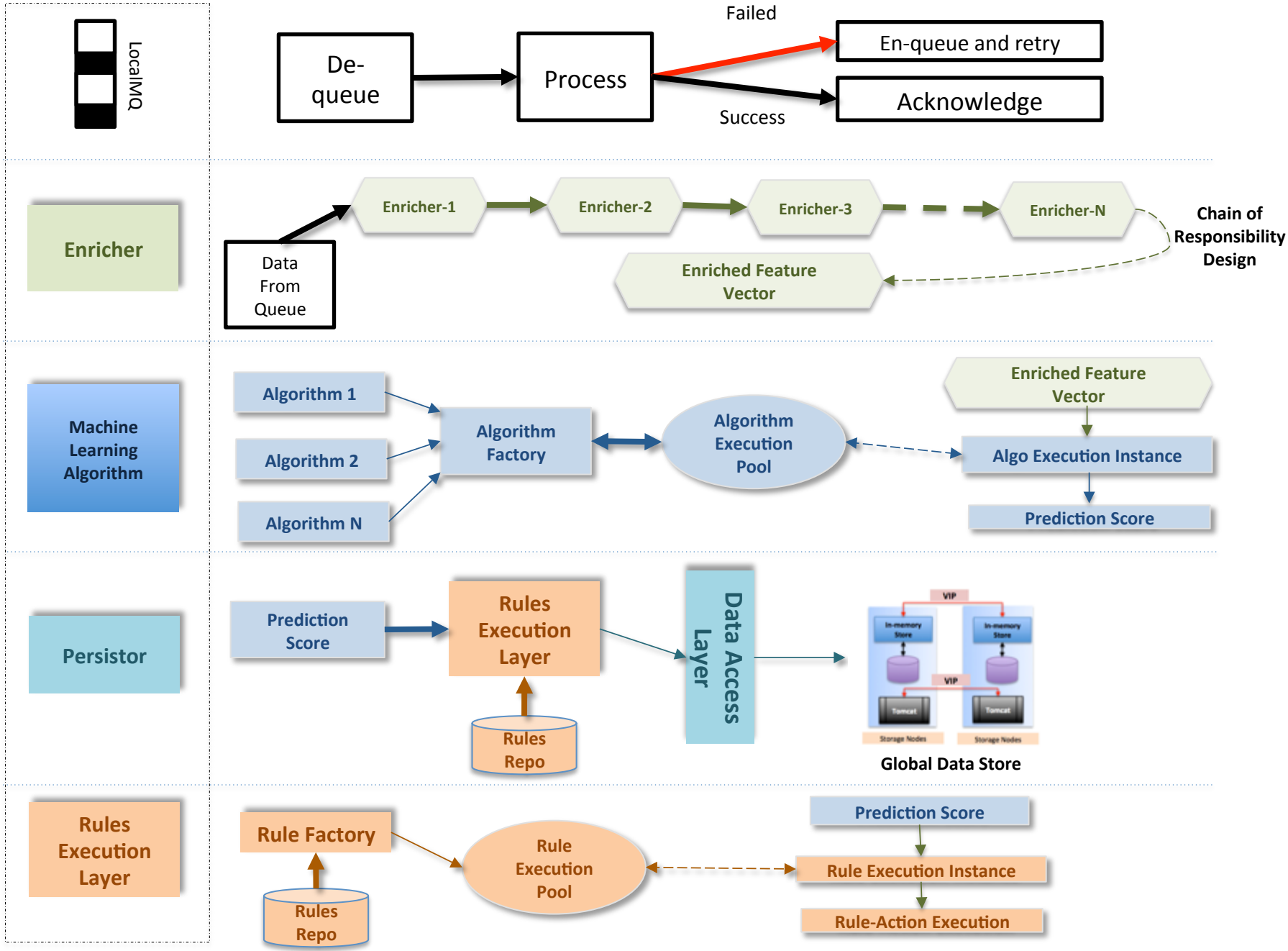
Big Data – Real Time Analytics Platform (CAPSTONE) : Physical Architecture



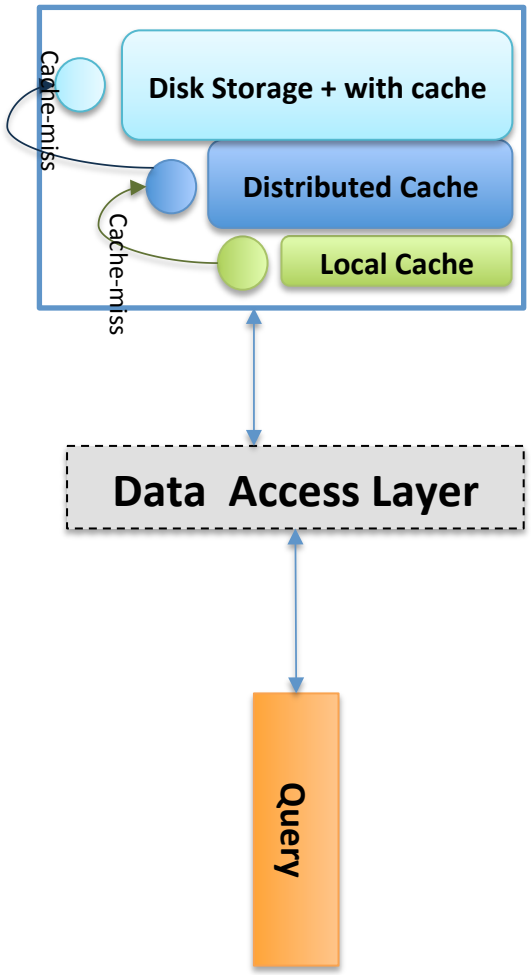
CAPSTONE – Elastic Computational Engine Architecture

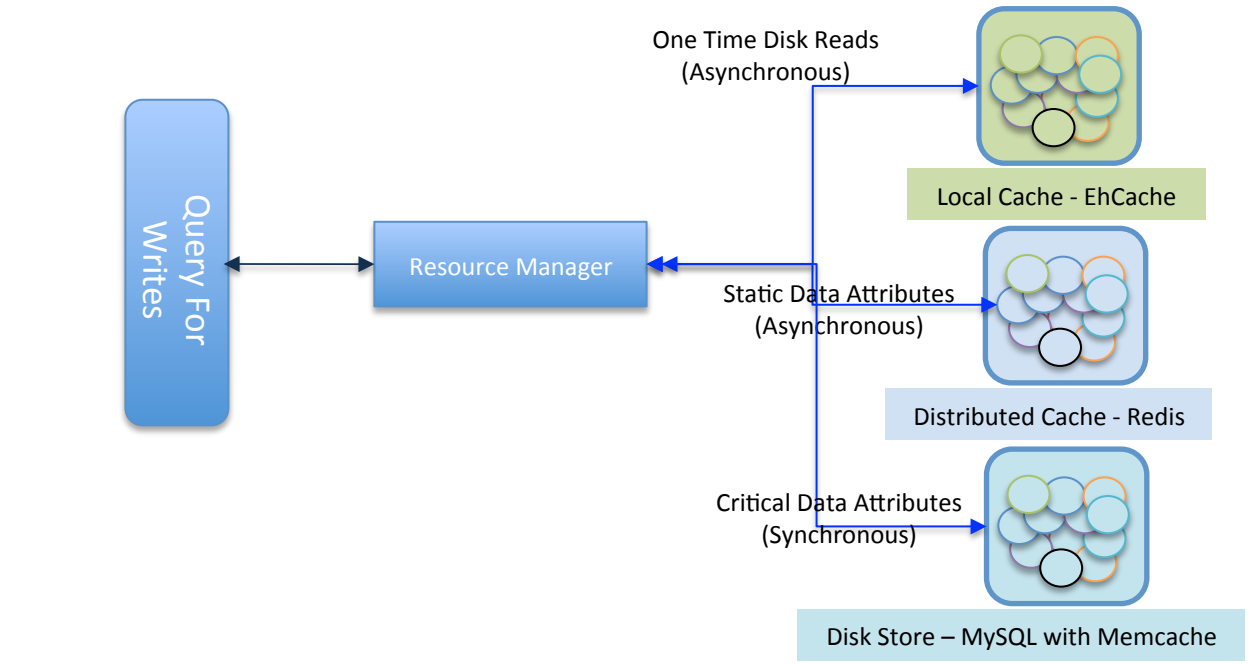


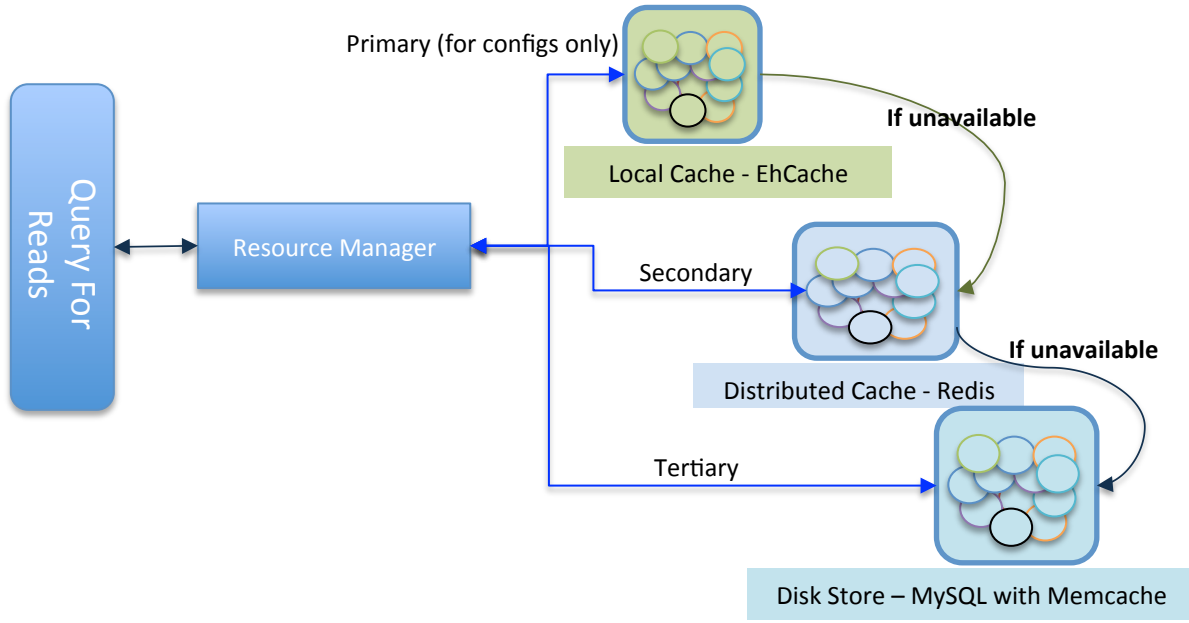
CAPSTONE – Elastic Computational Engine Architecture



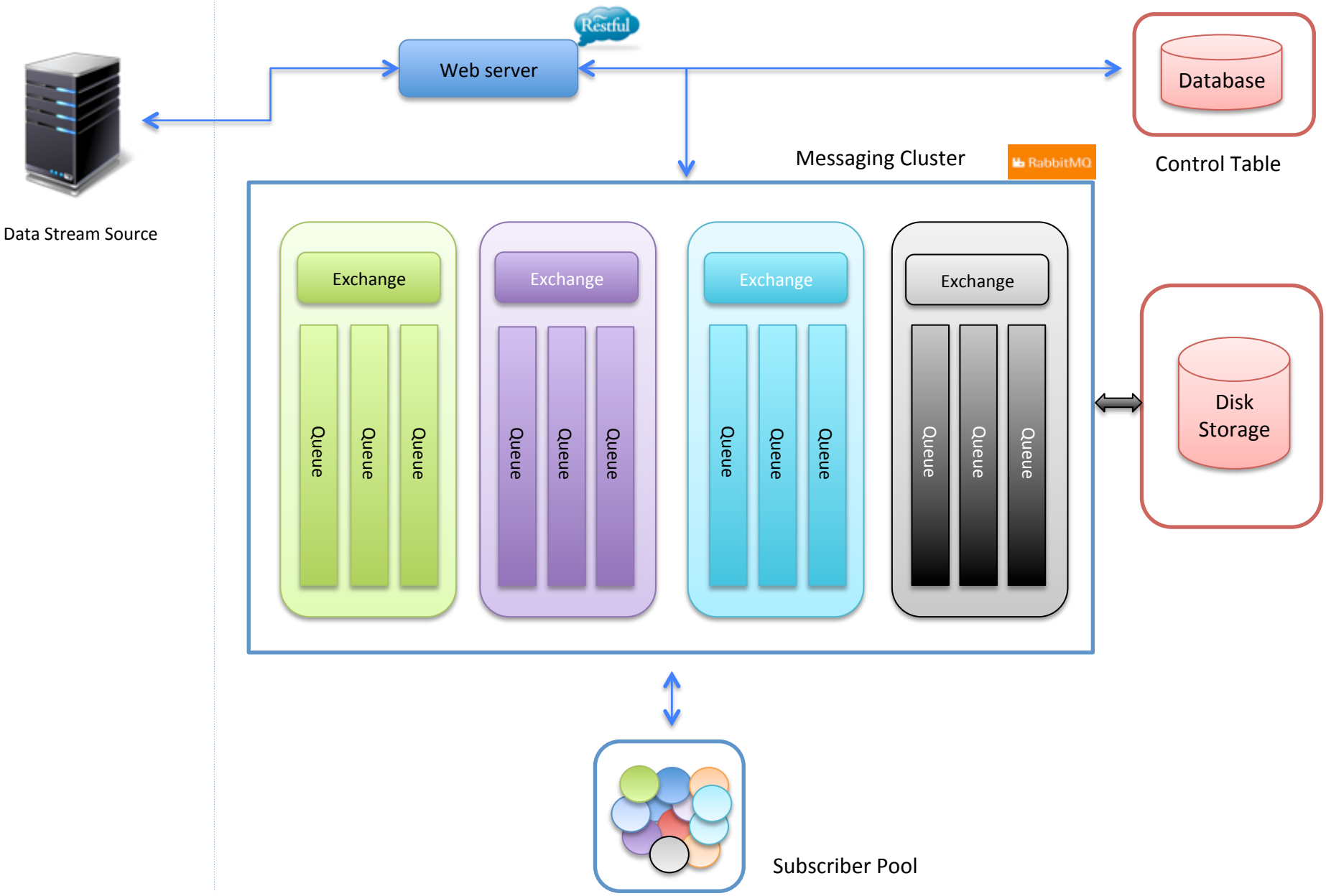
CAPSTONE – Architecting Caching Solution for Realtime







CAPSTONE – Messaging Cluster Architecture



CAPSTONE – Self Service Layer Architecture

Databases

In-Memory Store

Query Engine

Data Access Layer

Security, Authentication and Authorization Layer

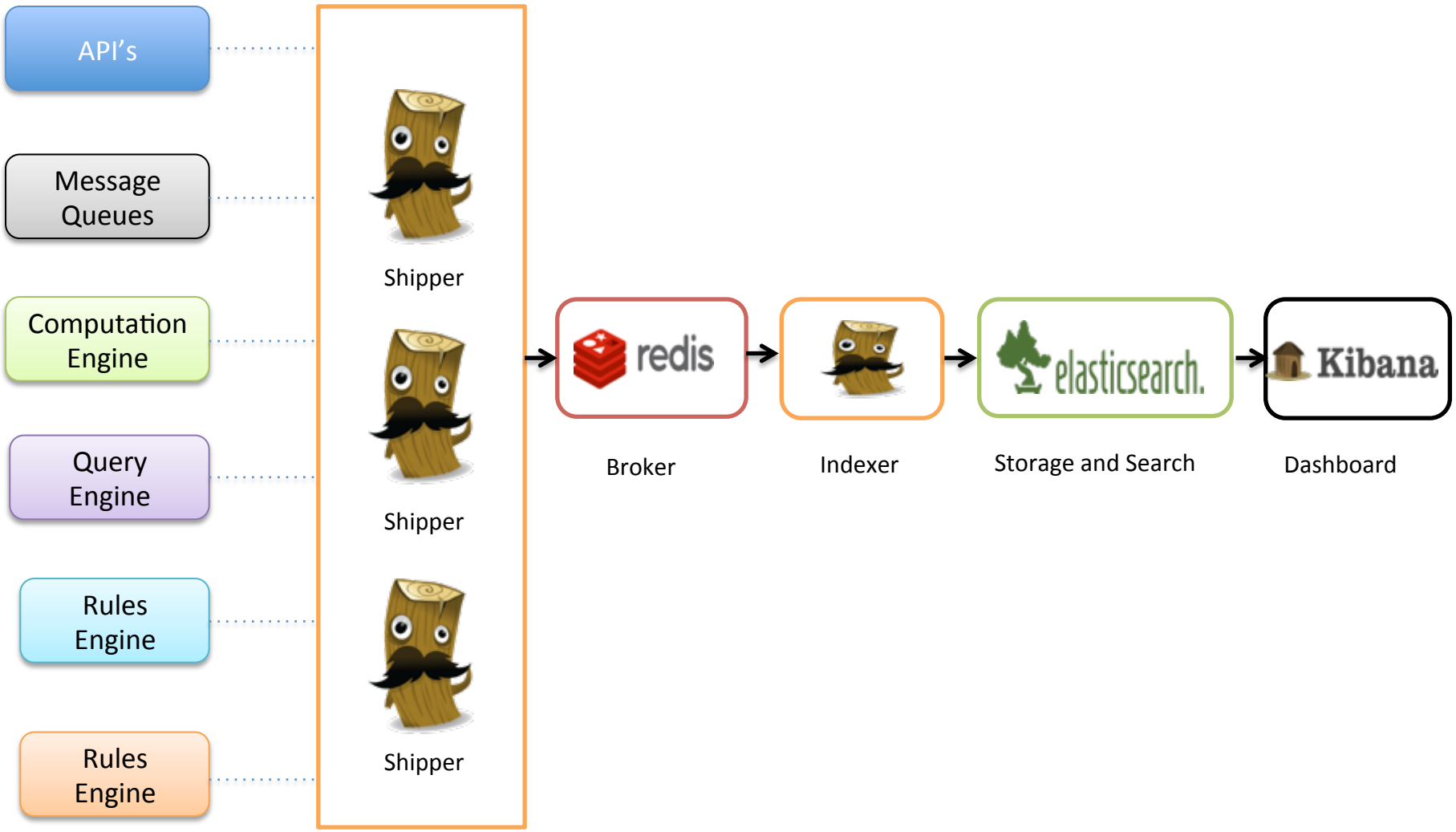
API Layer

Visualization and Reporting Layer

Performance and System Monitoring Layer

External Application Interface

CAPSTONE – Logging reporting layer



CAPSTONE – Summary of prototype