JUBATUS & Predictive Analytics and Sensing – Deep Dive - Draft**

October 7, 2015

NTT DATA

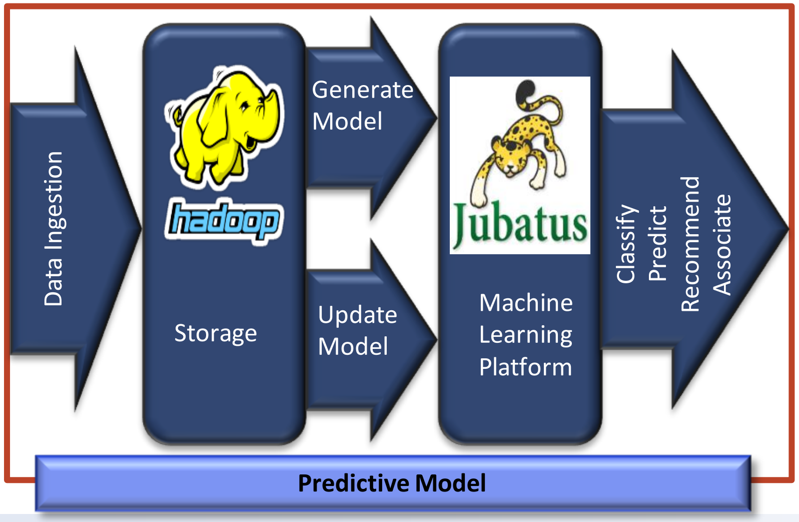
Predictive Analytics and Sensing

NTT DATA's Big Data Analytics Toolkit provides an advanced Machine Learning framework and platform – **Jubatus**
Jubatus is a processing platform for real-time analysis of flow-type data, capable of supporting large volumes within a distributed, scalable architecture that achieves massive performance.

BENEFITS

- Powerful and robust techniques for predictive analysis
- Informed decision making leveraging predictive and scoring models
- Support for various machine learning modules like – Classifier, Regression, Recommender, Anomaly Detection, Graph Mining
- Improved cost efficiency and profitability with predictive analytics
- Data preprocess and feature extraction
- Full range of feature conversion functions (from unstructured data to ML formats)

DELIVERY



Client Experience: Successful application of Machine Learning technology in all industry sectors, enabling use cases that advance well beyond traditional BI to achieve continuous improvement in prediction.

	Average Project	Large Project
Scenarios:	Intelligent decision making, call routing	
Timelines:	4 -12 weeks	12+ weeks
Team:	Lead Data Scientist, 1-3 Data Scientists	Lead Data Scientist, 4-7 Data Scientists
Deliverables:	Computing framework for real-time analysis of big data, Future state architecture and recommendations, production analytics	
Results:	Learning, prediction, recommender models, prescription	

DIFFERENTIATORS

Distributed Online Machine Learning Framework
Enables fixed time computation, high scale, fault tolerance

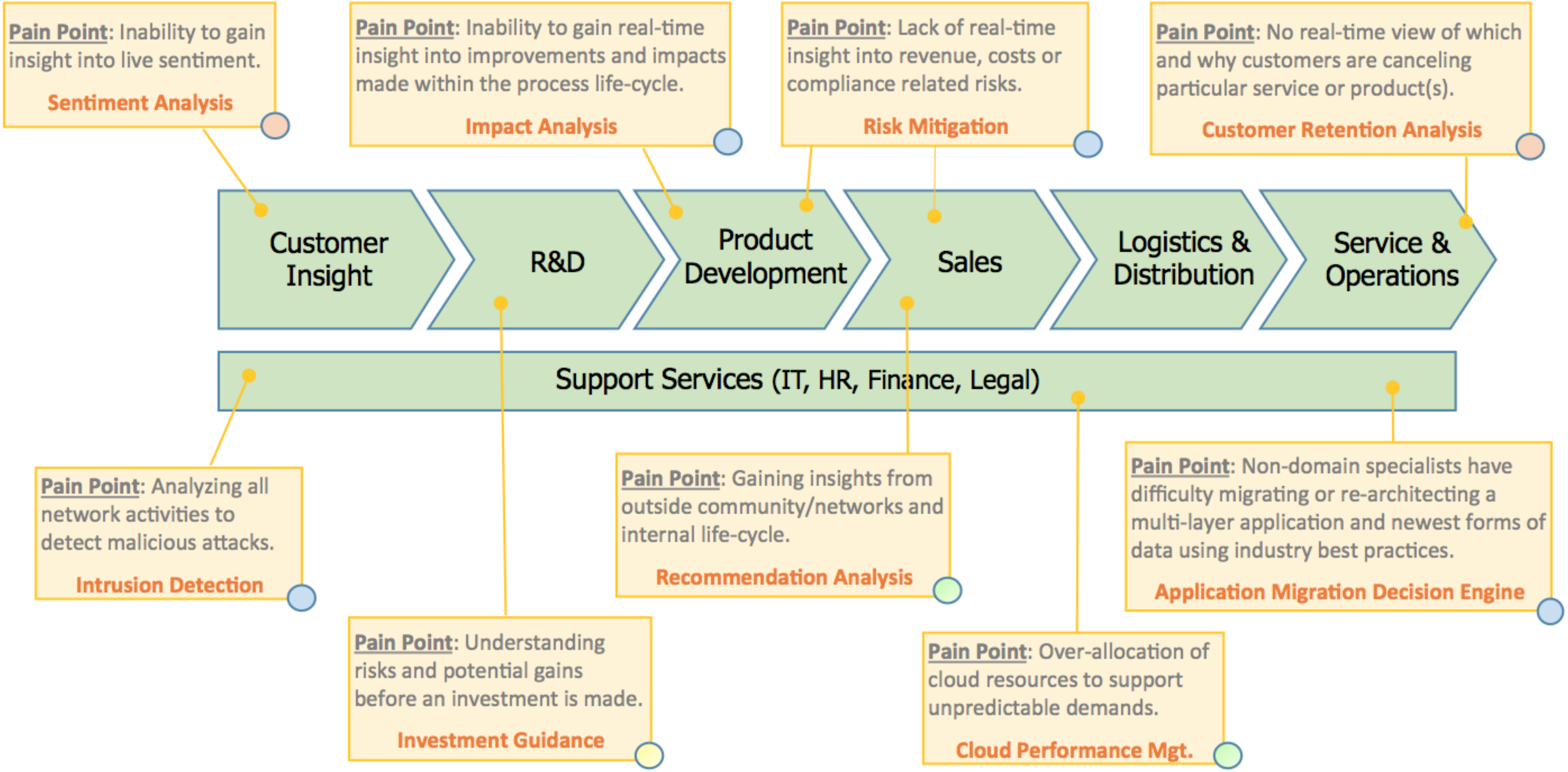
Stateful Stream Processing
"Push-type" enabling continuous sensing and learning of arriving data

Synchronization Framework
Ability to perform 'training' and 'results sharing' in parallel.

Analysis Templates

b

Our out of the box analysis templates *accelerate the development and implementation* of analytical models, speed up the pattern discovery, increases accuracy of business problem definition, and include **best practices from *200 Big Data implementations* .**



Pain Point: Inability to gain insight into live sentiment.

Sentiment Analysis

Pain Point: Inability to gain real-time insight into improvements and impacts made within the process life-cycle.

Impact Analysis

Pain Point: Lack of real-time insight into revenue, costs or compliance related risks.

Risk Mitigation

Pain Point: No real-time view of which and why customers are canceling particular service or product(s).

Customer Retention Analysis

Pain Point: Analyzing all network activities to detect malicious attacks.

Intrusion Detection

Pain Point: Gaining insights from outside community/networks and internal life-cycle.

Recommendation Analysis

Pain Point: Non-domain specialists have difficulty migrating or re-architecting a multi-layer application and newest forms of data using industry best practices.

Application Migration Decision Engine

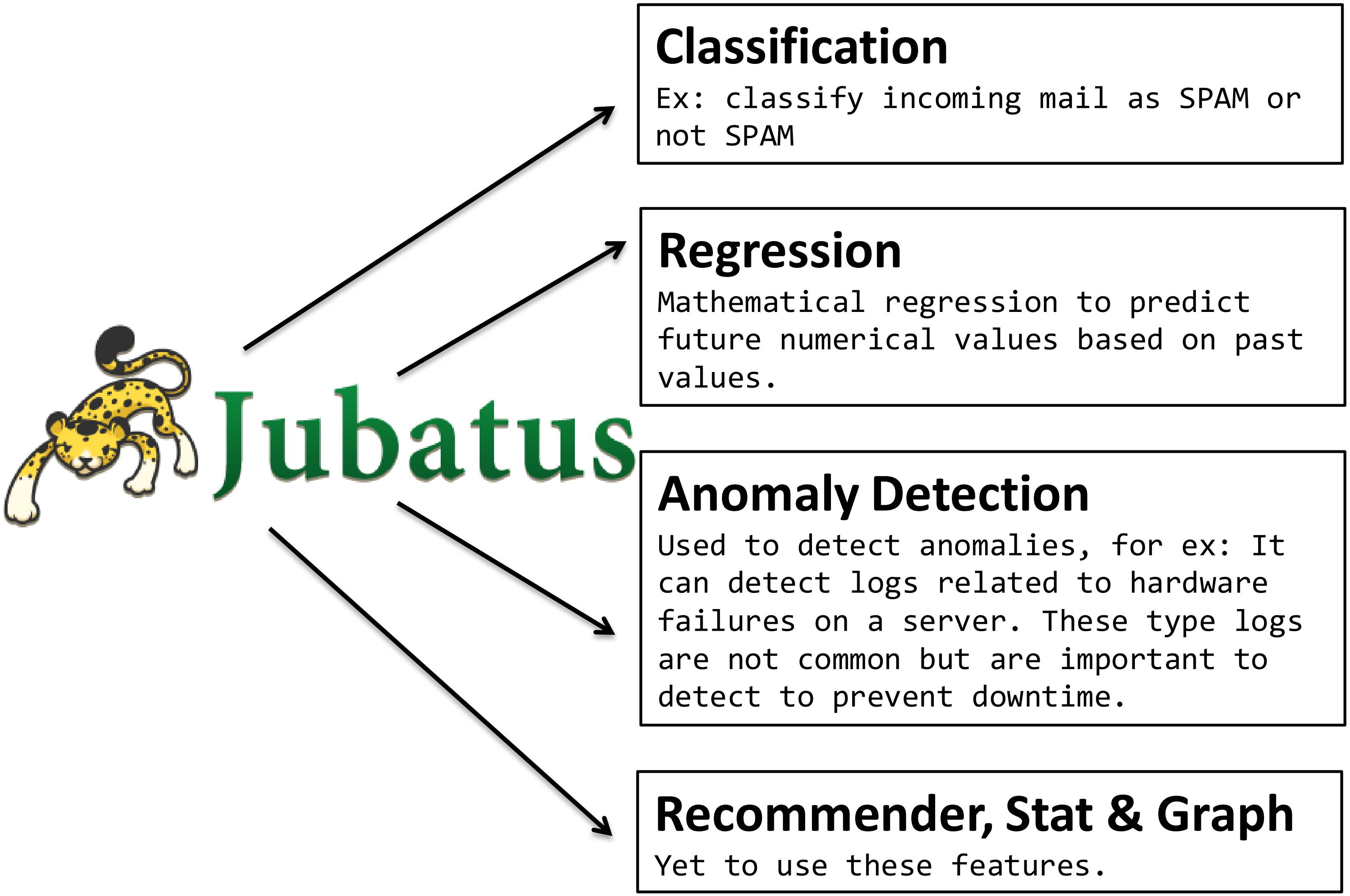
Pain Point: Understanding risks and potential gains before an investment is made.

Investment Guidance

Pain Point: Over-allocation of cloud resources to support unpredictable demands.

Cloud Performance Mgt.

Online Machine Learning Capability



Classification

Ex: classify incoming mail as SPAM or not SPAM

Regression

Mathematical regression to predict future numerical values based on past values.

Anomaly Detection

Used to detect anomalies, for ex: It can detect logs related to hardware failures on a server. These type logs are not common but are important to detect to prevent downtime.

Recommender, Stat & Graph

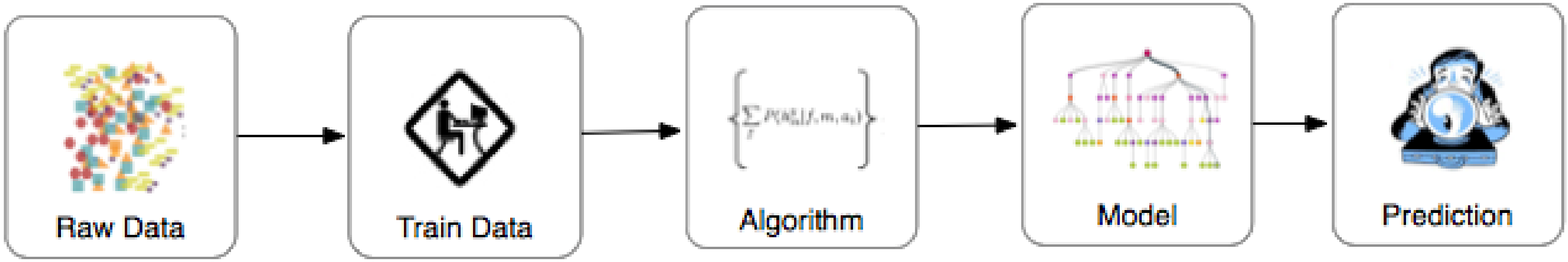
Yet to use these features.

Machine Learning Templates

Machine Learning Templates

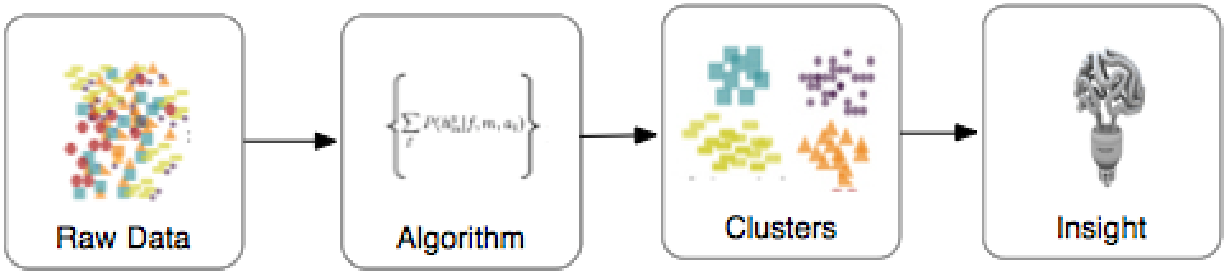
Supervised learning: From raw data to prediction

Machine learning that employs a training dataset as the basis for predictive analysis.



Unsupervised learning: From raw data to pattern detection

Analysis of unlabeled data for the purpose of finding patterns, clusters, outliers, etc.



Real-time Machine Learning Modules

Below is a listing of algorithms available in the NTT accelerator suite.

Real-time algorithm catalogue:

Approach	Algorithm	Parameters
Classification (supervised)	Perception	None
	Passive Aggressive	Regularization weight
	Confidence Weighted	Regularization weight
	Adaptive Regularization of Weight Vectors	Regularization weight
	Normal Herd	Regularization weight
Clustering (unsupervised)	K-means	K, method, bucket size/len, bicriteria base, forgetting factor/threshold
	Gaussian Mixture	
Regression (supervised)	Passive Aggressive	Sensitivity, Regularization weight
Anomaly (unsupervised)	Local Outlier Factor	(reverse) Nearest neighbor num

Batch-mode Machine Learning Modules

C

Batch algorithm catalogue:

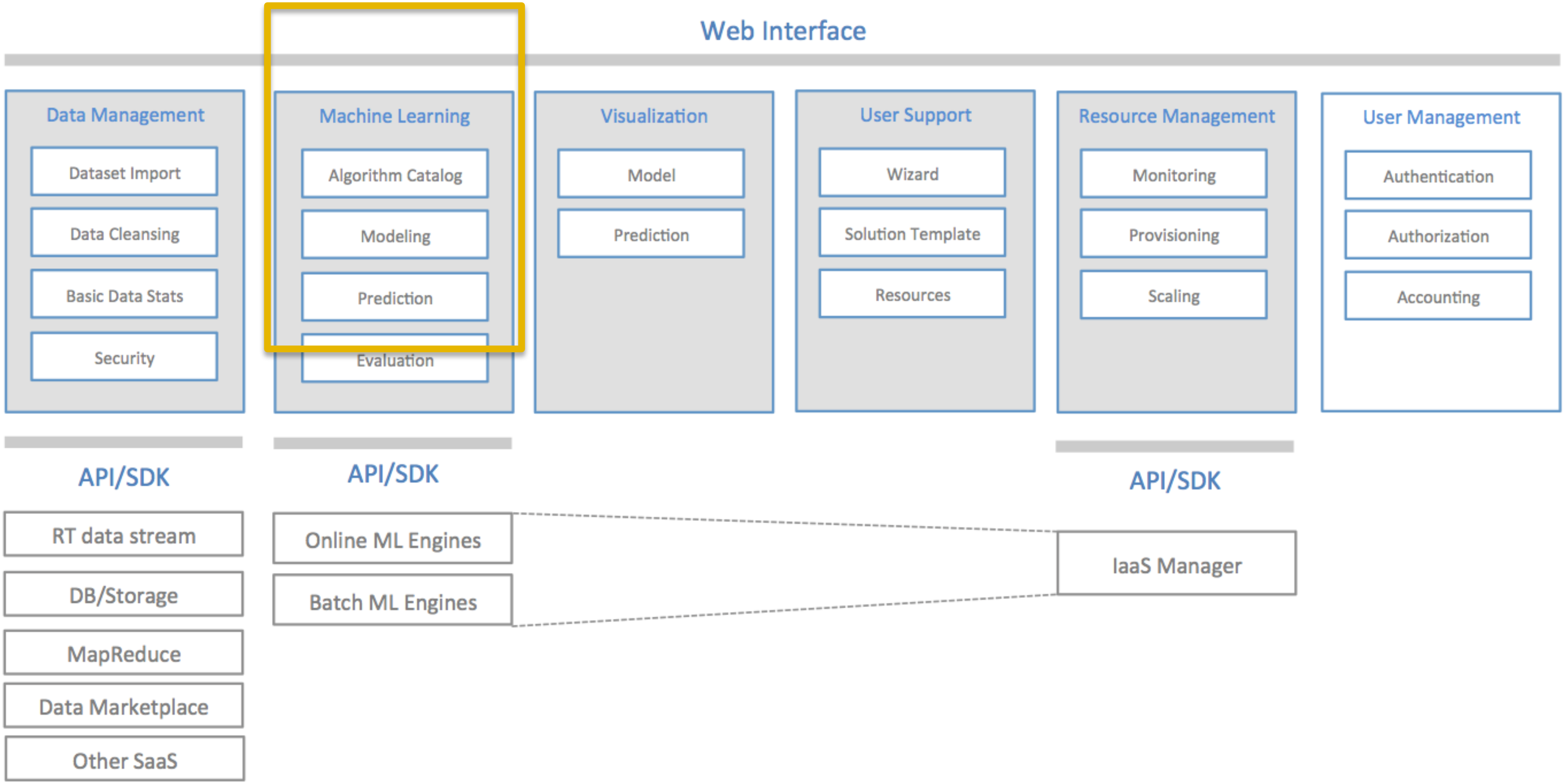
Approach	Algorithm	Parameters
Classification (supervised)	Support Vector Machine (SVM)	C, gamma, kernel
	Kernel approximation	N_components
	KNeighbors Classifier	K (nearest neighbours), Weights
	SVC Ensemble	N_estimators, max_features
	Naïve Bayes	Alpha, class_prior, fit_prior
	Random Forest	N_estimators, max_features
	Decision Trees	Max_depth, max_features
Clustering (unsupervised)	MeanShift	Bandwidth, seeds
	KMeans	N_clusters, max_iter, n_jobs, n_init
	Spectral Clustering GMM	n_components, covariance_type, random_state, n_iter, n_init
Regression (supervised)	SGD Regressor	Loss, penalty, alpha, l1_ratio
	ElasticNet Lasso	Alpha, l1_ratio, fit_intercept, precompute
	Support Vector Regression (SVR)	C, gamma, kernel
	SVR Ensemble	C, gamma, kernel
Ensemble (supervised)	Random Forests	n_estimators, max_features
	AdaBoost	n_estimators, learning_rate
	Gradient Tree Boosting	(various)

Example Useage by Domain

Use-Case Scenario	Short description
Customer Retention Analysis	> Predict which users are more likely than average to cancel a particular service or return a product. Allows operators to act proactively make process changes.
Contents/ product recommendation	> Recommend new media contents or products to users based on their previous ratings
Risk Mitigation	> Predict loan defaults or late payments in order to manage risk
Many more	> Intrusion detection, Impact analysis, investment guidance…

Component View

Predictive Analytics, and Big Data Capability Components View



Web Interface

Data Management
- Dataset Import
- Data Cleansing
- Basic Data Stats
- Security

Machine Learning
- Algorithm Catalog
- Modeling
- Prediction
- Evaluation

Visualization
- Model
- Prediction

User Support
- Wizard
- Solution Template
- Resources

Resource Management
- Monitoring
- Provisioning
- Scaling

User Management
- Authentication
- Authorization
- Accounting

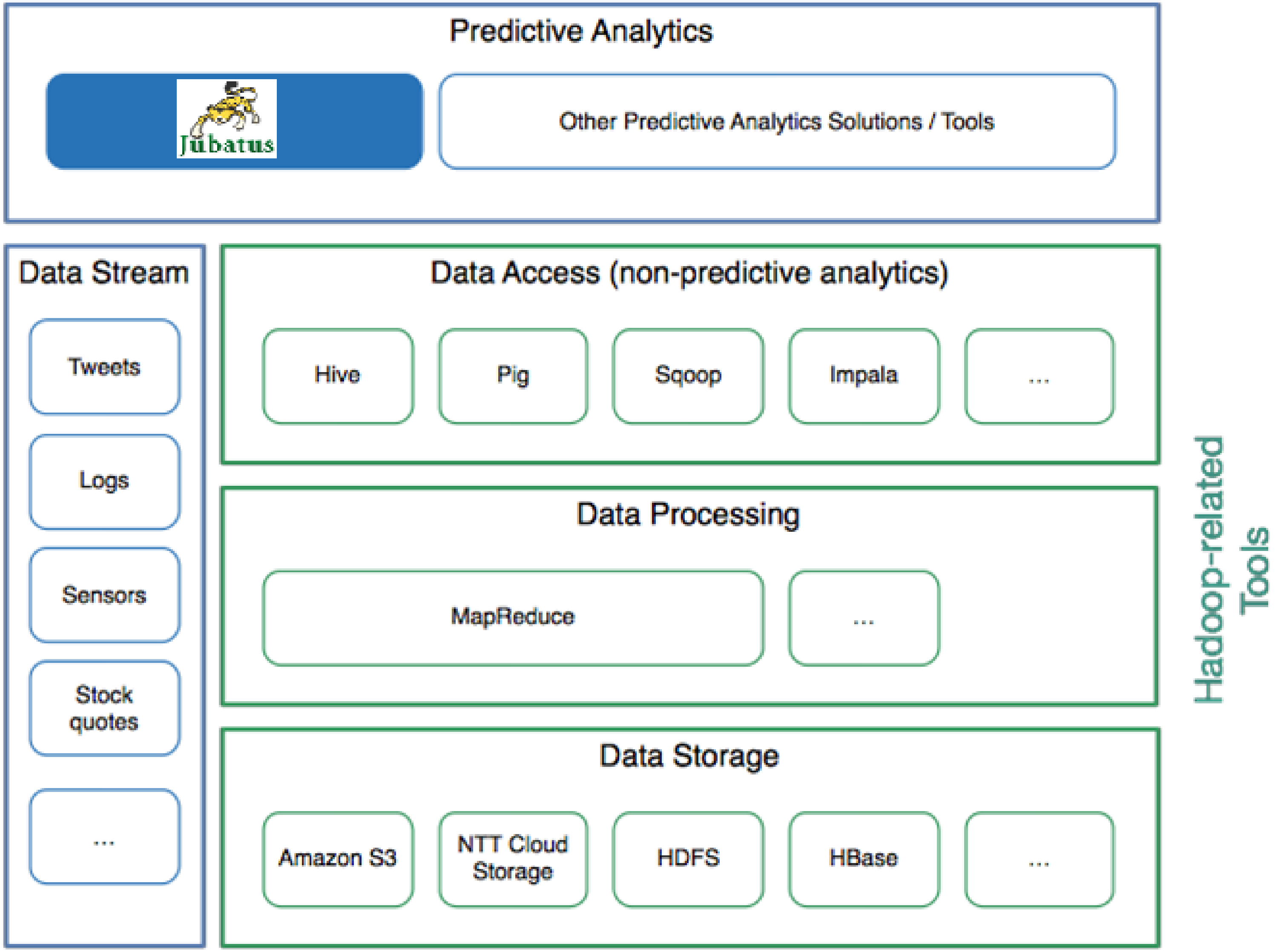
API/SDK
- RT data stream
- DB/Storage
- MapReduce
- Data Marketplace
- Other SaaS

API/SDK
- Online ML Engines
- Batch ML Engines

API/SDK
- IaaS Manager

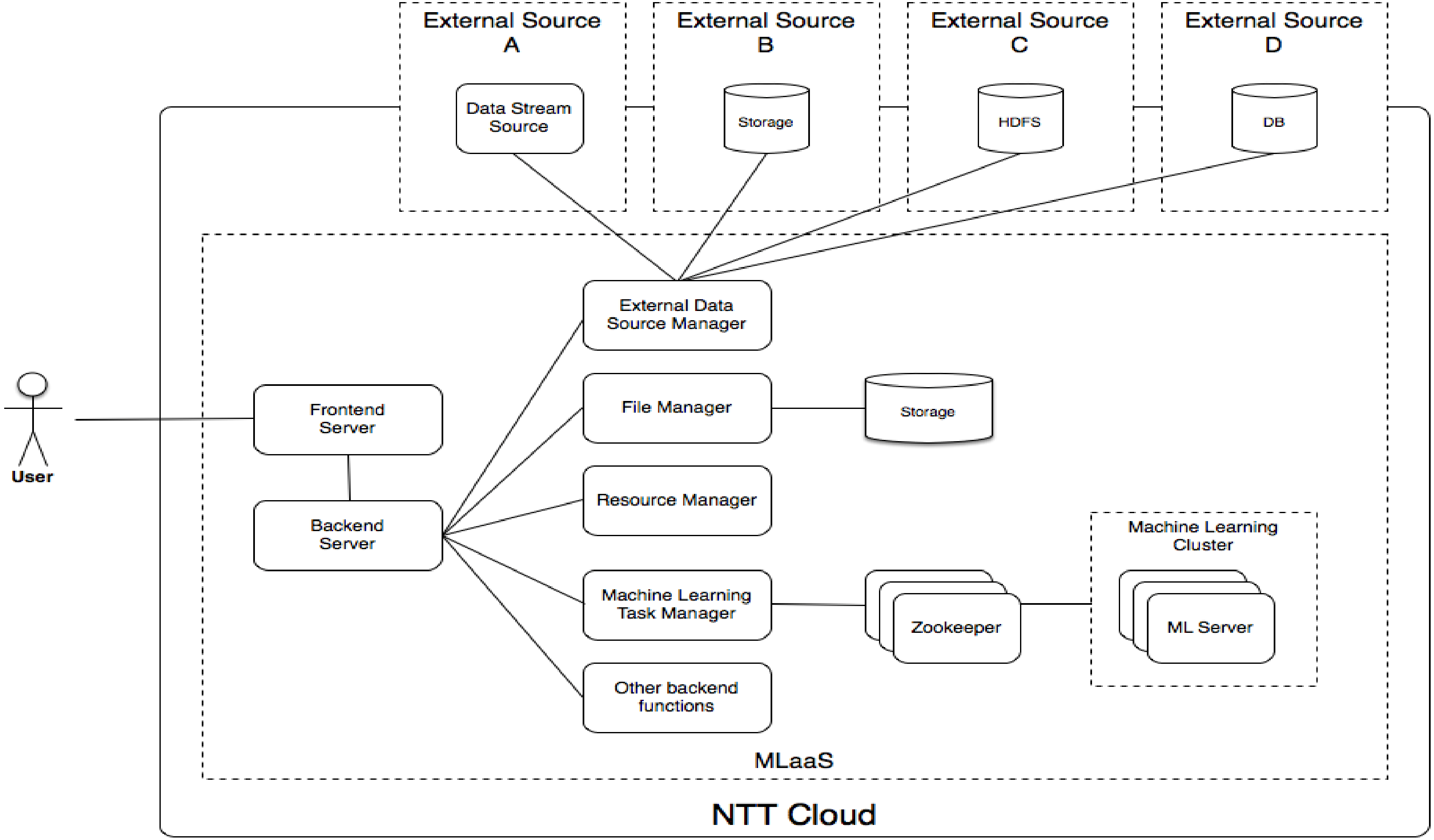
Heart of the system, performing the data analysis

Machine Learning & Predictive Analytics Components



Deployment View

d



**The deployment view will not change much if the final deployment is in a data center rather than a cloud service.

Big Data Accelerator Suite Benefits

3

- **Quick Prototyping and Implementation** of ICM solution
- **Accelerates Analysis process and model development** for ICM algorithms
- **Enhanced Accuracy** of Big Data use cases
- **Rapid Discovery** of metadata and patterns
- **Improved Quality and Clarity** of end-user information
- **Highly usable, intuitive** visualization
- **In-built elasticity** through de-coupling of data from the hardware
- **Low cost implementation** based on open-source software
- **Enablement** of a variety of user skill levels (basic to expert users)
- **Handling of all types of data** through predictive analytics components
- **Turnkey system:** service requires no development, deployment, provisioning, or maintenance
- **Integration** of multiple applications
- **Real-time** power
- **Flexibility and sustainability** of analytics through algorithm catalogue