Quixey Engineering Gap Analysis
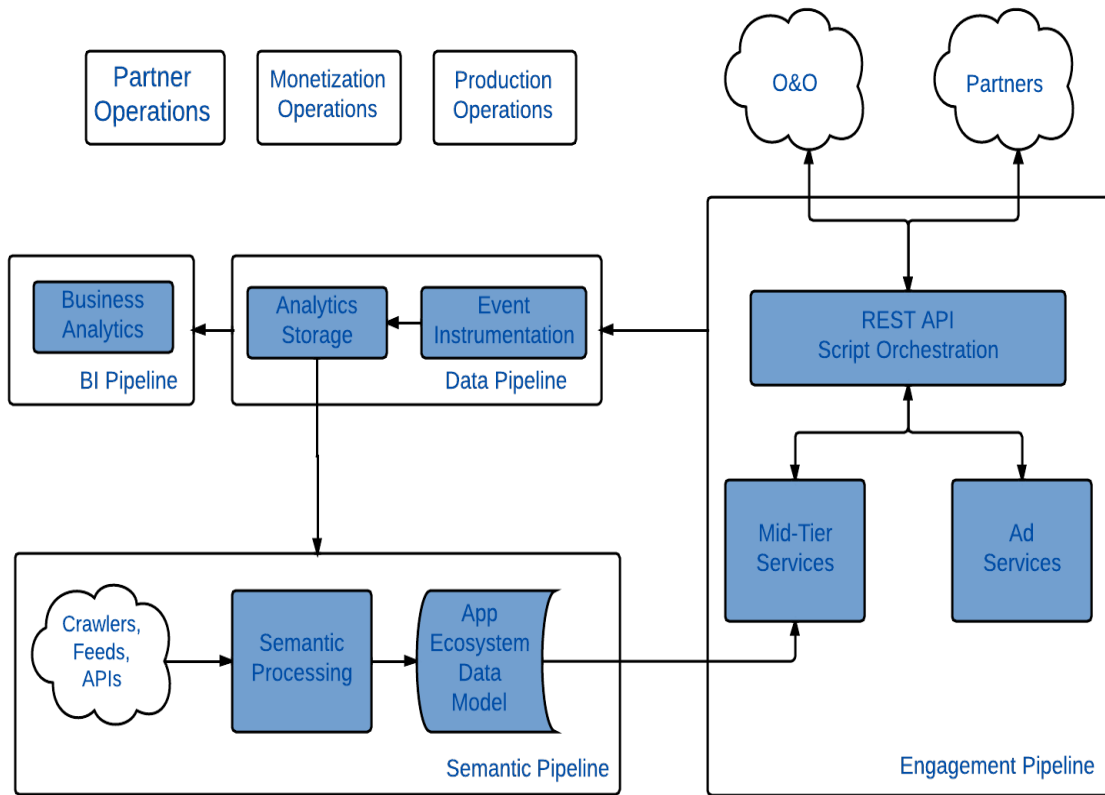
Alibaba Team Visit

Project Phoenix

Mountain View, California

January 21, 2014

Quixey 3.1 Architecture

Gap Analysis Topics

- Deploy Engagement Pipeline v3

Gap Analysis Topics: Engagement Pipeline

- Rest API / Script Orchestration:
  - Deploy Engagement Pipeline v3: In progress; expected completion mid-February
- New work: Separate service and data deployment code from engagement pipeline services to avoid developer involvement in IAAS differences
  - Metadata Mid-Tier Service (in progress)
  - Search Mid-Tier Service (not yet started)
  - Query Categorization Service (not yet started)
  - Develop uniform service and data deployment DevOps deployment processes and scripts:
    - Mid-tier service code and data changes
    - Front-end service code and AAA changes
    - Consider Container-based deployment to allow EP artifacts to be deployed multi-cloud
      - AWS creates instances from AMI and deletes them; pre-paids are accounting only
      - Ali Cloud uses pre-purchased instances only and servers must be re-furbished
      - Containers allow separation of developer/prodops, and separation of instance create/destroy vs re-use
- KVS/Cache systems:
  - Isolate Redis/Memcache/ECS usage and replace with OCS or custom-deployed Redis/Memcache
  - Engagement Pipeline v3 Metadata Service Cache Population design issue: target fix March 1
- Additional work still necessary on Quixey side for Mid-Tier Services not proposed for M1/M2 to avoid forks (move to v3 pipeline):
  - Autosuggest Service (in progress)
  - Spelling (not yet started)
  - Trending (not yet started)
  - Match service (split into Metadata Query and Sim matrix; not yet started)

Gap Analysis Topics: Ads

- Engagement Pipeline

  - Ads is not required for M1, but if required for M2 or M3 the Ad pipeline and Ad Operations systems need to be ported.

  - Ads Mid-Tier service is REST/SOA based and uses same components as other Quixey pipelines

- Monetization Operations

  - Ads Analytics pipeline uses same components as other Quixey pipelines (Kafka, Hadoop, RDS)

  - Campaign Manager: Aimed at inside sales, but may still need Chinese UI

  - Partner Portal: Not easily ported to new systems and so needs to be analyzed

Gap Analysis Topics: Semantic Pipeline (Hadoop)

- For M1/M2, deploy on a single-node Hadoop instance for 300K apps

  - Reduced risk and improved TTM to avoid cluster management and OSS/Hadoop integration

  - Requires Hadoop usage optimization in edition merging to optimize two $O(N^2)$ steps

- Complete generation of search Avro index (in progress)

  - Features to generate are a moving target due to search relevance changes for Chinese language

  - Integrate language analysis in feature generation

- Complete automation with Luigi and Jenkins (in progress)

- Develop testable assertions for Chinese app market ("Angry Birds" tests)

  - For incoming data sets

  - For pipeline step results

Gap Analysis Topics: Data Pipeline (Hadoop/Kafka)

- Deploy Kafka infrastructure in M1

- Use single-node Cloudera Hadoop instance to consume Kafka events to OSS

- Perform analytics reporting at Quixey for test & development

- Deploy BI pipeline as M2 approaches (next slide)

    - Eliminate need for multi-node Hadoop cluster and RDS in M1 if there is no BI and no Ads

Gap Analysis Topics: BI Pipeline (Hadoop)

- Add BI Pipeline as M2 approaches

  - Hadoop at scale required for large amounts of traffic (cluster management)

  - New BI reports likely necessary for new market

  - Language/localization issues for reports and data

Gap Analysis: Production Operations

- Addressed separately

Action Items from Workshop

Component	Gaps Identified	Resolution	Who	When
?	?	?	(lead), names	