Data Ontology Cache System

The Monadics! · TWO SIGMA

Engineers: Derek Mansen, Marc Weil, Matt Kotsenas, Robbie Gladmon, Tom Rudick
Faculty Advisor: Dr. J. Scott Hawker

Project Sponsor: Mark Roth, Two Sigma Investments LLC.
Senior Project 2010

Software Engineering
Rochester Institute of Technology

Company Background

- Process-driven investment manager with billions in assets under management
- Offices in New York, Houston, and London
- Invent trading algorithms that utilize petabytes of financial data
- Employ state-of-the-art technologies such as event stream processing and grid computing
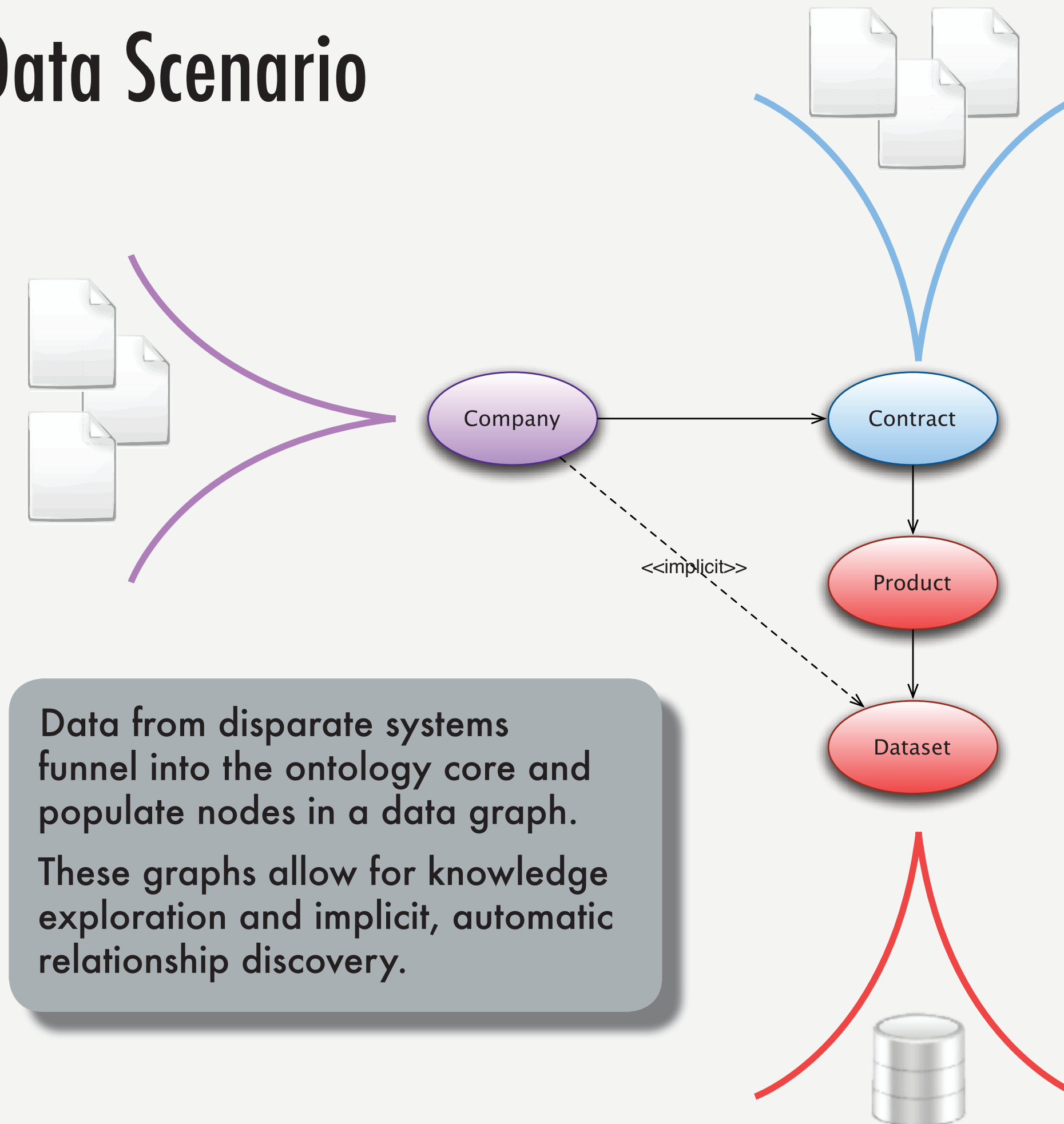
Project Motivation

- Many different teams produce artifacts stored on disparate, unconnected systems
- Currently difficult to ask questions about related data that span multiple systems
- Unable to sanity check and normalize data for storage on specific volumes
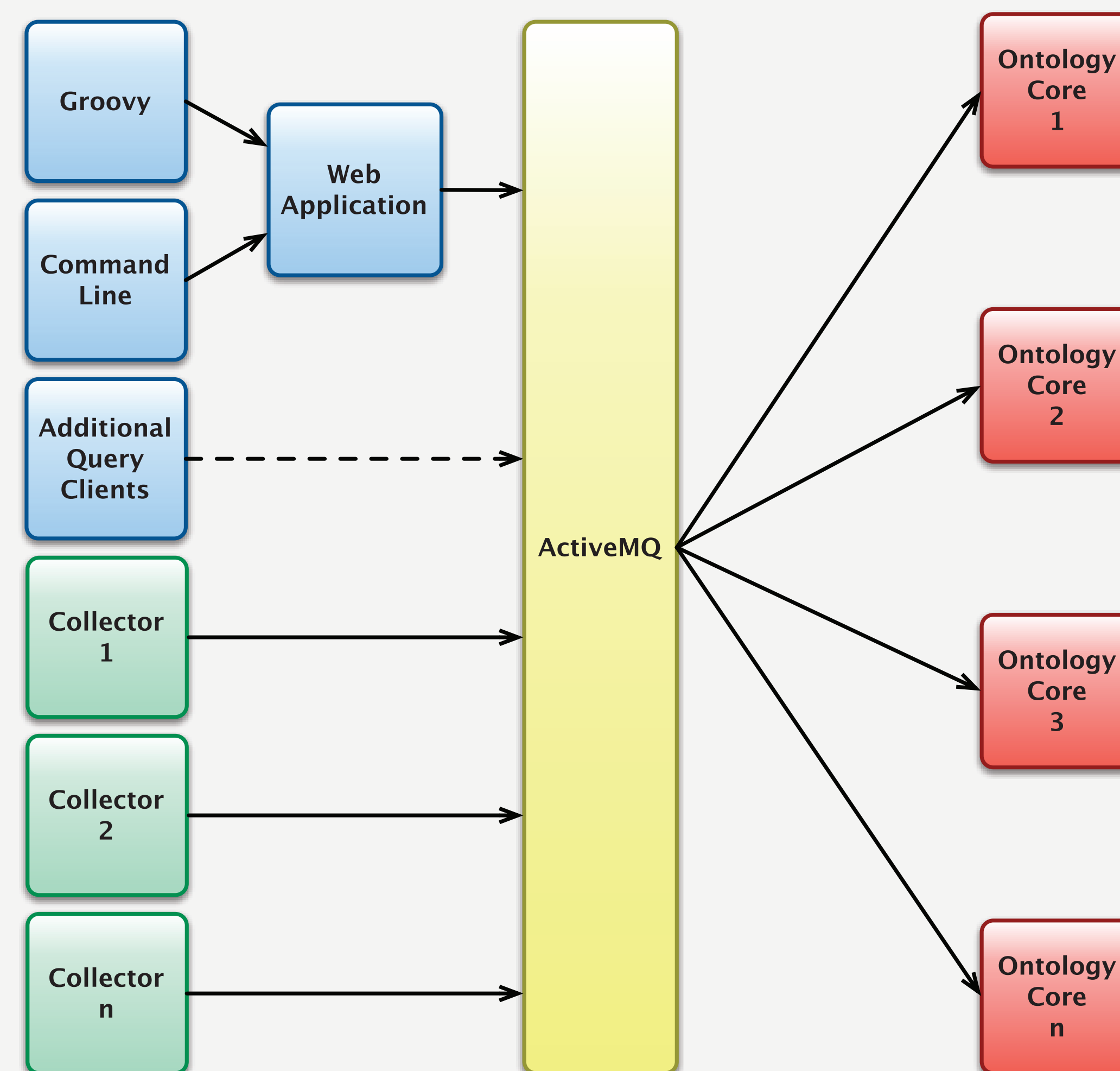
Technologies

Semantic Web · ActiveMQ · Groovy · Java · Tomcat · Jena semantic web framework · spring

Data Scenario

Company → Contract
<<implicit>>
Product
Dataset

Data from disparate systems funnel into the ontology core and populate nodes in a data graph.

These graphs allow for knowledge exploration and implicit, automatic relationship discovery.

System Architecture

Groovy, Command Line → Web Application
Additional Query Clients
Collector 1
Collector 2
Collector n
→ ActiveMQ →
Ontology Core 1
Ontology Core 2
Ontology Core 3
Ontology Core n

System Features Delivered

- Distributed ontology cores
  - Security built into the ontology
  - Accepts data pushes from collectors
  - Responds to SPARQL queries from clients
  - Data validation against ontology schema
- Web application
  - Relationship discovery with graph visualizer
  - Full-text search engine
  - Groovy reporting module
- System status monitor
- UNIX-style command line tools
- Java API for query clients and collectors

Dual-Model Process

- First Quarter: Evolutionary Prototype
  - Technology spikes and feasibility studies
  - Minimize risk
- Second Quarter: Time-Boxed Iterations
  - Implementation of features
  - Maximize value

Evolutionary Prototyping · Timeboxed Development
Risk · Value
Technologies Tech Spike · Core Technologies Tech Spike · Web Technologies Tech Spike · Prototype 1 · Prototype 2 · Iteration 1 · Iteration 2 · Iteration 3 · Iteration 4