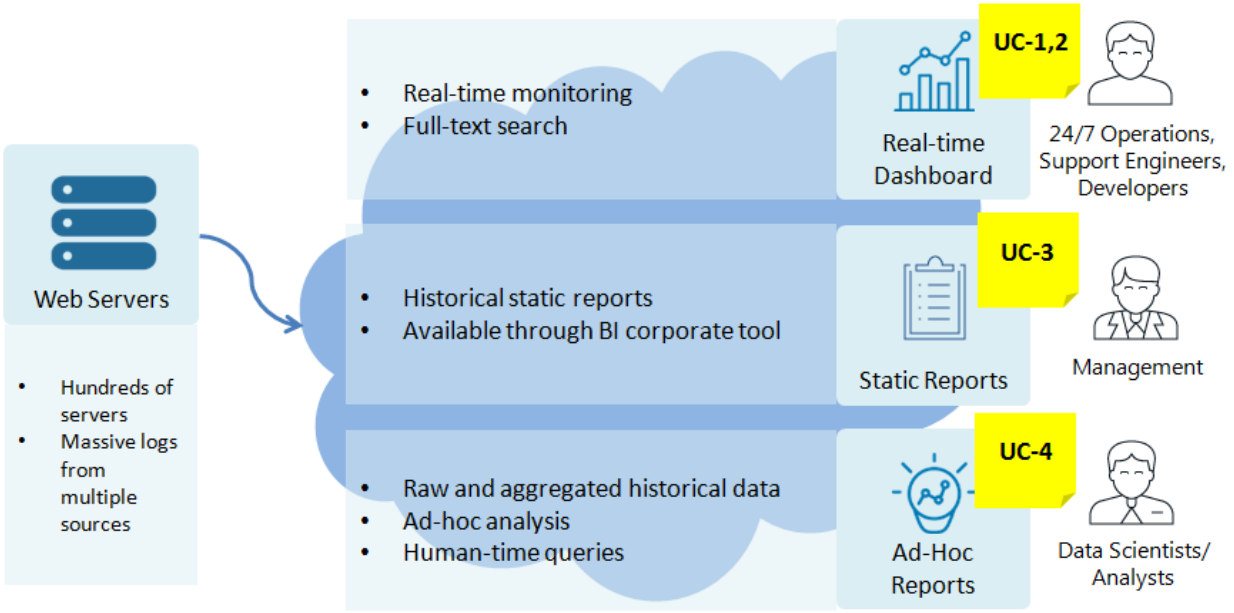
Smart Decisions:
Game Scenario - Big Data

You are a software architect who is designing the architecture for an important Big Data system which involves the collection and analysis of a massive set of semi-structured data coming from hundreds of servers.

The customer is an Internet company that provides popular content to millions of web users. The company's IT department realizes that existing legacy monitoring systems cannot handle anymore the required capacity as well as requests from company stakeholders including the Operations Team, Data Scientists and Management who would like to leverage all the various data that can be collected.
Finally, the main project requirements are agreed, the budget approved and you, as a seasoned architect, are called to design the new ambitious project. You are provided with the following marketecture diagram:

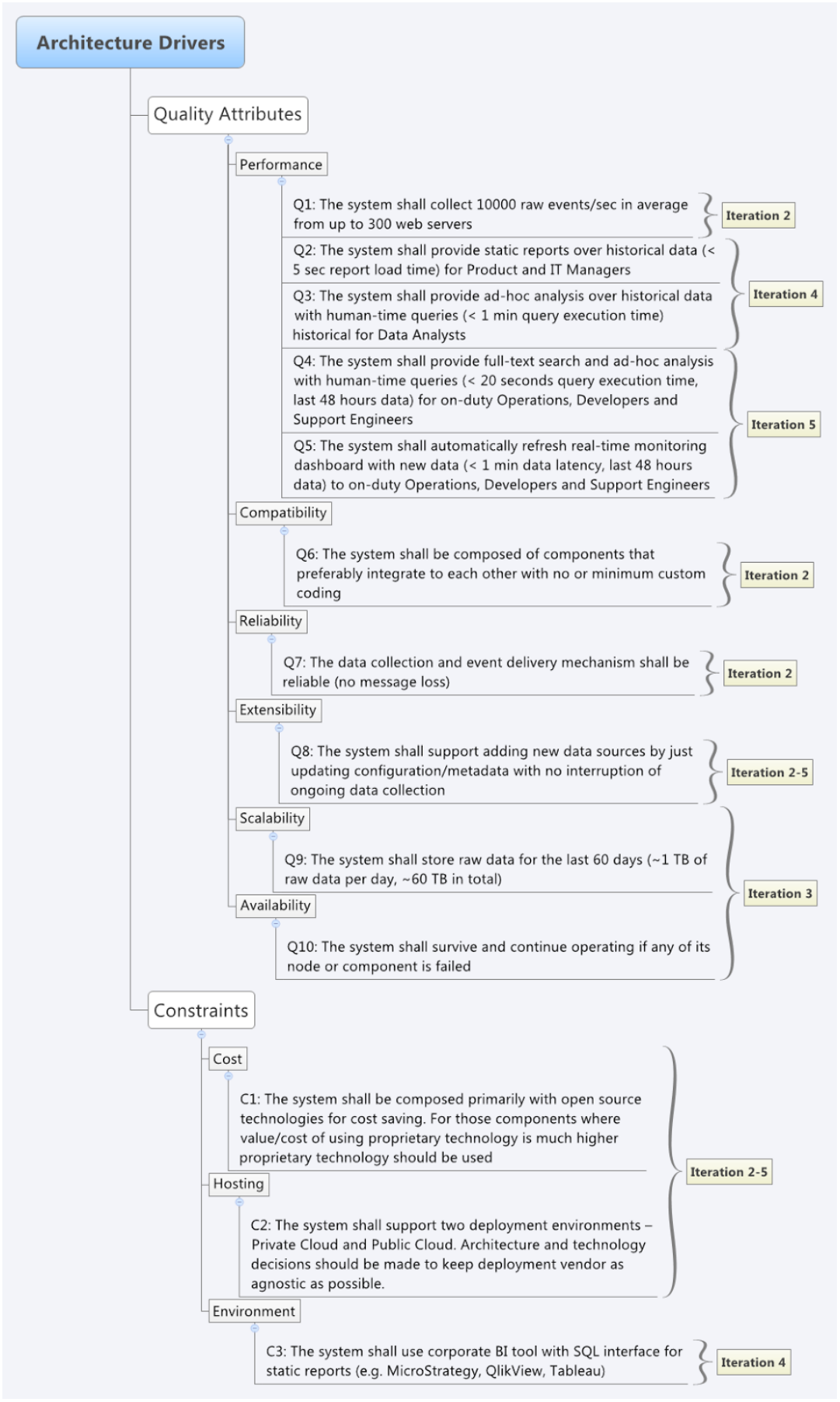


The primary use cases for the system are the following:

Use case name	Description
UC-1: Monitor online services	On-duty Operations Team can monitor the actual state of services and IT infrastructure (such as web servers load, user activity and errors) through real-time operational dashboard to react on potential issues ASAP

UC-2: Troubleshoot online service issues	Operations Team (including Support Engineers and Developers) can do troubleshooting and root cause analysis on the latest collected logs by searching the log patterns and filtering messages
UC-3: Provide management reports	Managers, including Product and IT Managers, can see historical information through predefined (static) reports in a corporate BI tool, such as system load in time, features usage, SLA violations and quality of releases
UC-4: Provide ad-hoc data analytics	Data Scientists and Analysts can do ad-hoc data analysis through SQL-like queries to find out specific data patterns and correlations to improve infrastructure capacity planning and customer satisfaction

The quality attributes and constraints for the system are the following:


Architecture Drivers
Quality Attributes
Performance
Q1: The system shall collect 10000 raw events/sec in average from up to 300 web servers
Iteration 2
Q2: The system shall provide static reports over historical data (< 5 sec report load time) for Product and IT Managers
Q3: The system shall provide ad-hoc analysis over historical data with human-time queries (< 1 min query execution time) historical for Data Analysts
Iteration 4
Q4: The system shall provide full-text search and ad-hoc analysis with human-time queries (< 20 seconds query execution time, last 48 hours data) for on-duty Operations, Developers and Support Engineers
Q5: The system shall automatically refresh real-time monitoring dashboard with new data (< 1 min data latency, last 48 hours data) to on-duty Operations, Developers and Support Engineers
Iteration 5
Compatibility
Q6: The system shall be composed of components that preferably integrate to each other with no or minimum custom coding
Iteration 2
Reliability
Q7: The data collection and event delivery mechanism shall be reliable (no message loss)
Iteration 2
Extensibility
Q8: The system shall support adding new data sources by just updating configuration/metadata with no interruption of ongoing data collection
Iteration 2-5
Scalability
Q9: The system shall store raw data for the last 60 days (~1 TB of raw data per day, ~60 TB in total)
Availability
Q10: The system shall survive and continue operating if any of its node or component is failed
Iteration 3
Constraints
Cost
C1: The system shall be composed primarily with open source technologies for cost saving. For those components where value/cost of using proprietary technology is much higher proprietary technology should be used
Hosting
C2: The system shall support two deployment environments – Private Cloud and Public Cloud. Architecture and technology decisions should be made to keep deployment vendor as agnostic as possible.
Iteration 2-5
Environment
C3: The system shall use corporate BI tool with SQL interface for static reports (e.g. MicroStrategy, QlikView, Tableau)
Iteration 4