Deployment and Operations for Software Engineers 2nd Ed

Chapter 6—Measurement

Outline

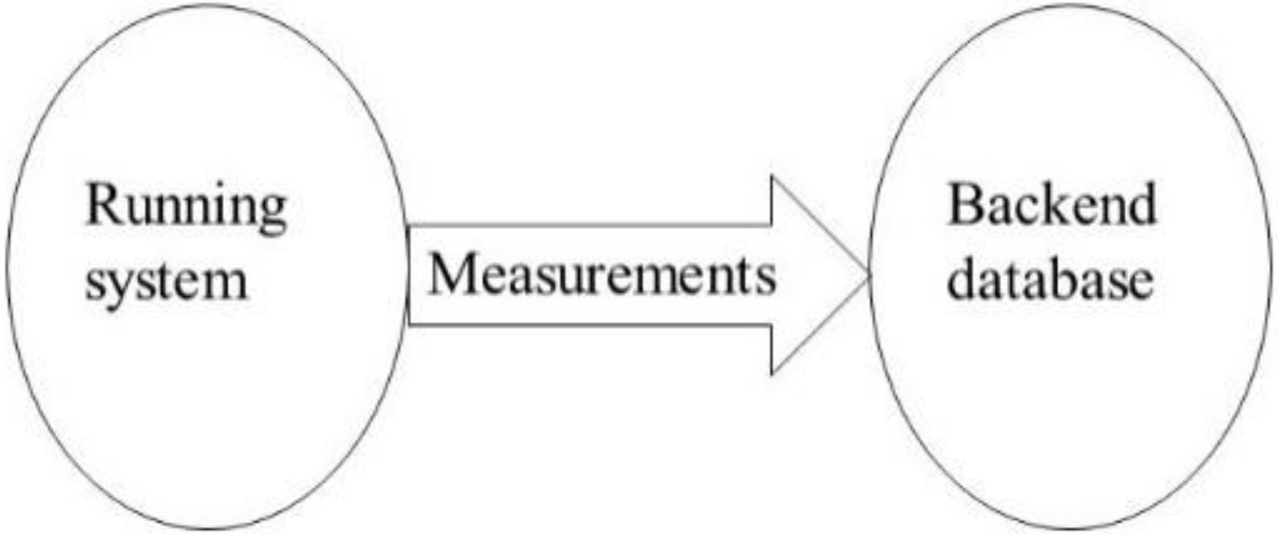
- **Overview**
- Logs
- Metrics
- Tracing

Purposes of gathering operational data

- Generate alerts. Measurements are used to generate alerts to indicate a serious problem with a system in operation.
- Forensics. When a problem has occurred, it must be identified. Logs collected by the system in response to events are used to identify the problem.
- Performance. Understanding the end-to-end behavior of a system in response to requests shows where the system spends time.

Measurement architecture

- Information is collected in a variety of locations within your system.
- It should be moved to a central location for analysis.
- This is typically a time series database.



Sources of measurement data

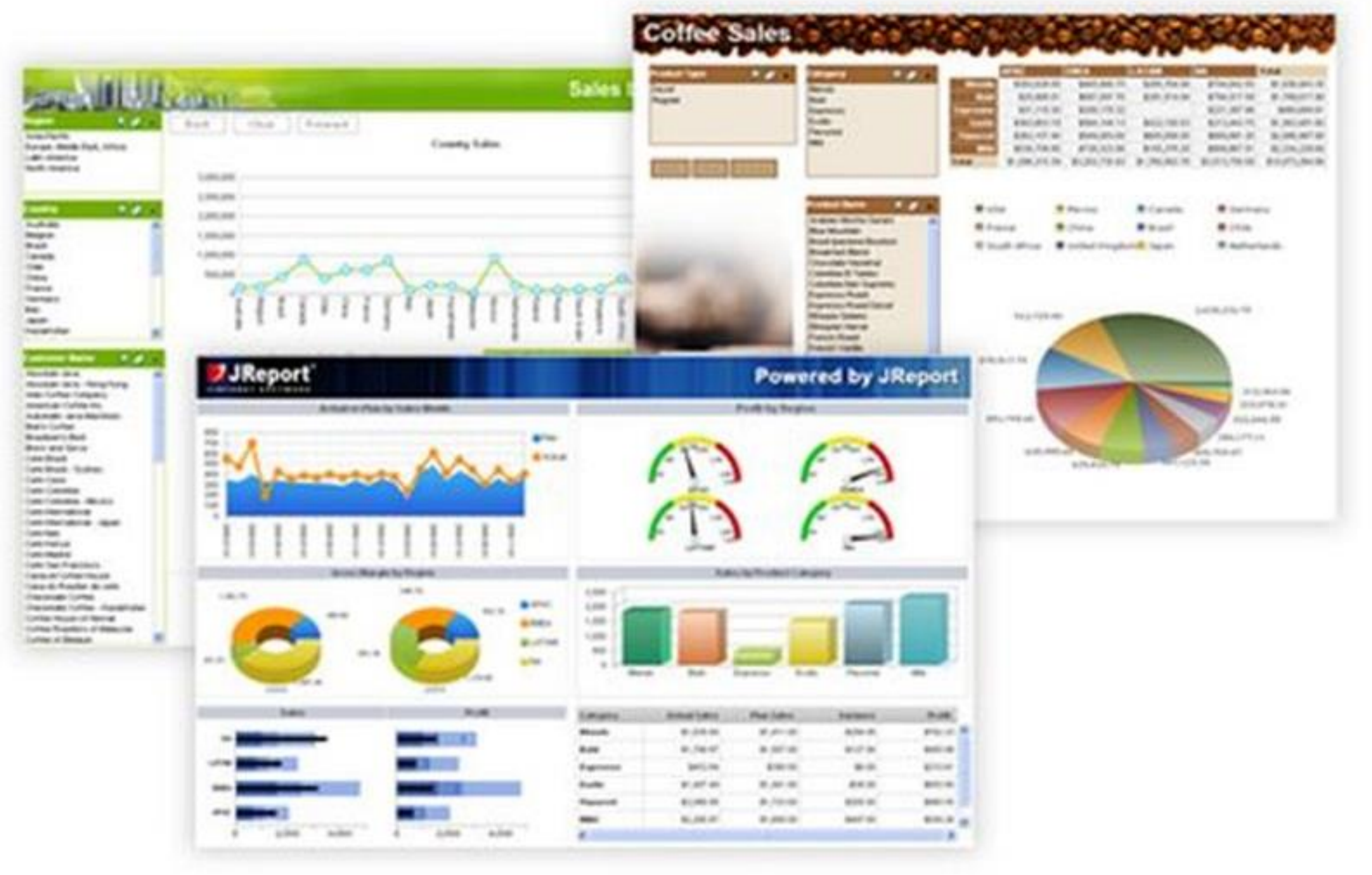
- The measurement data comes from two sources:
  - The infrastructure that monitors utilization,
  - The logs produced by the services.
- In both cases, there is a backend specific service that
  - gathers the information,
  - Reformats it for the backend,
  - passes it to the backend.

©Len Bass and John Klein 2022

Activities of backend

- The backend sends an alert based on a collection of rules similar to autoscaling rules.
- Displays a dashboard that allows an analyst to quickly determine whether there is a problem.
- Supports various types of analysis of the data collected.

Sample dashboard



Time readings in a distributed system

- Clocks on computers drift. ~1 second every 12 days.
- Clock measurement based on a single computer will be accurate to within clock resolution.
- Clock measurements taken across multiple computers cannot be used to reliably determine latency or sequence of events.

Time accuracy

- Network time protocol is accurate to within 1 millisecond on a local network and within 10 milliseconds on a public network.
- Compare these times to the time taken for messages or memory references. They are much larger.
- Tracing uses request IDs to identify sequences of activities.
- GPS time is accurate to about 100 nanoseconds but requires access to satellites.
- Atomic clocks have no measurable drift but are very expensive.

Discussion questions

1. What are some specific functions a backend database provides. See Splunk or Logstash for examples.
2. How could you compute the difference in clock readings between your smart phone and your laptop? Remember that it takes time to send a message.

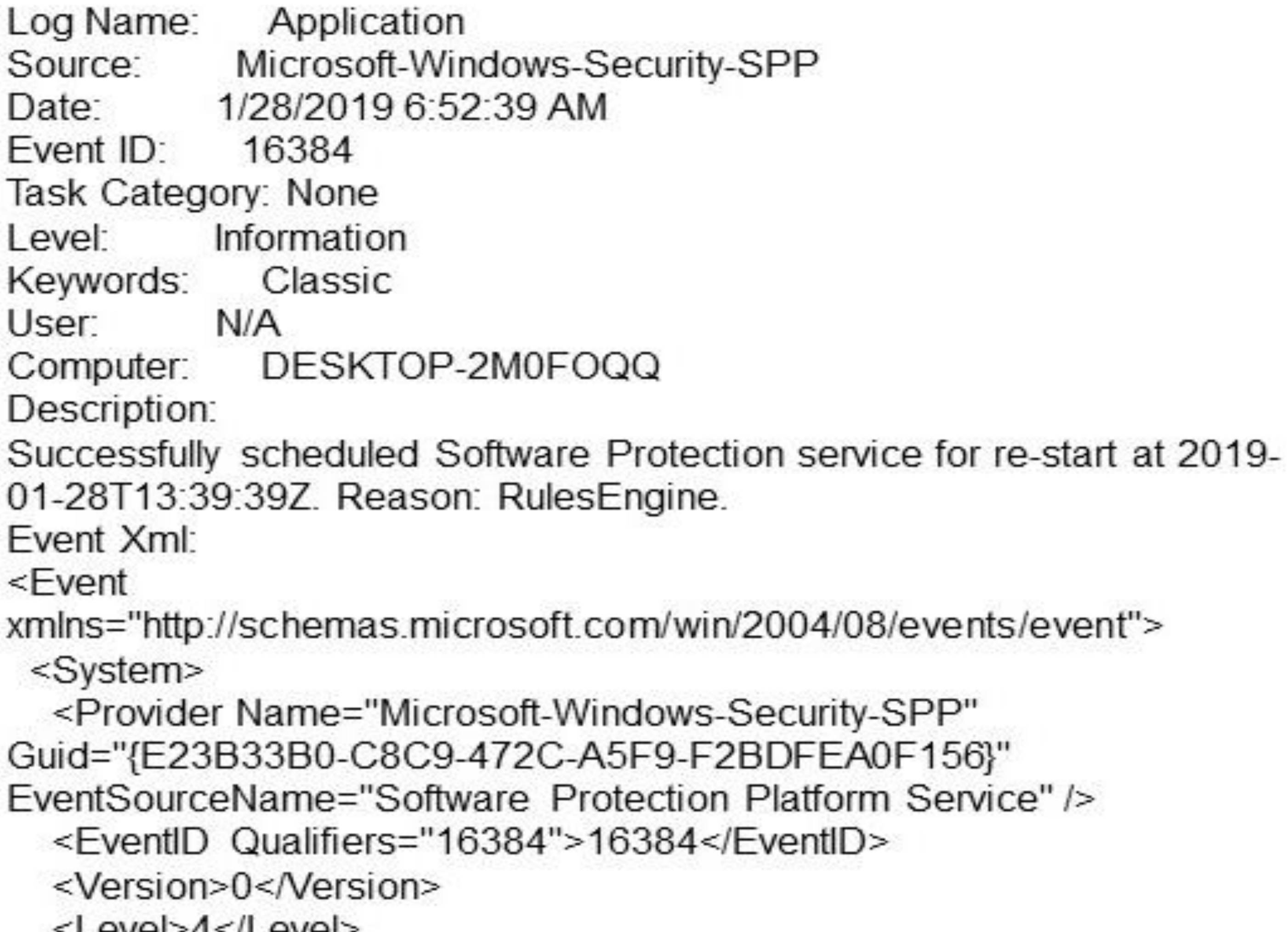
Outline

- Overview
- **Logs**
- Metrics
- Tracing

Log entry

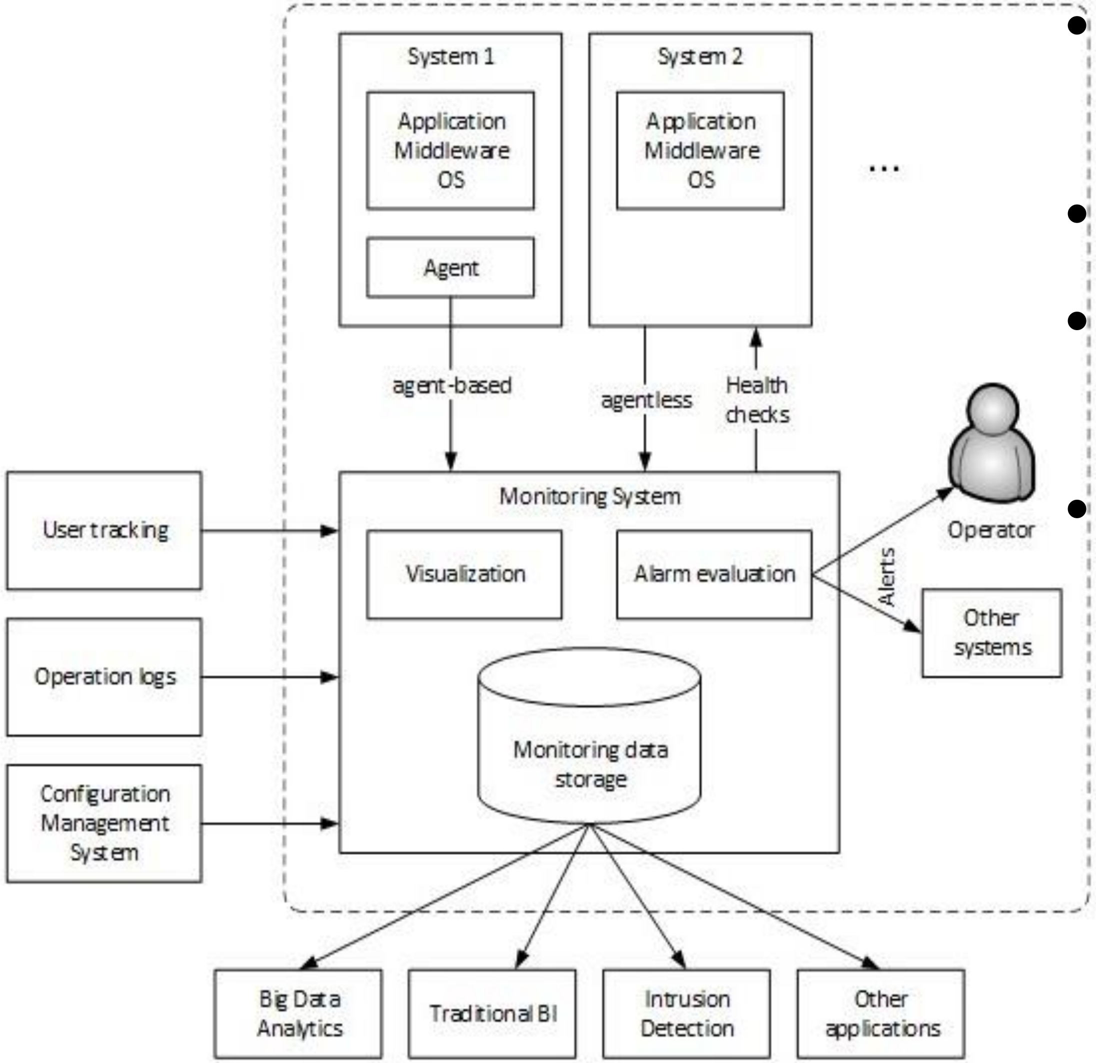
- A log entry is generated by a service in response to some event.
  - Entry/exit
  - Error detection
- Logs are used for forensic purposes. That is, finding the source of a problem.
- Contents of a log entry are typically
  - Time stamp
  - Identifying information
  - Reason for log entry
  - Variable values
  - …

Sample log from Windows



```
Log Name:      Application
Source:        Microsoft-Windows-Security-SPP
Date:          1/28/2019 6:52:39 AM
Event ID:      16384
Task Category: None
Level:         Information
Keywords:      Classic
User:          N/A
Computer:      DESKTOP-2M0FOQQ
Description:
Successfully scheduled Software Protection service for re-start at 2019-
01-28T13:39:39Z. Reason: RulesEngine.
Event Xml:
<Event
xmlns="http://schemas.microsoft.com/win/2004/08/events/event">
  <System>
    <Provider Name="Microsoft-Windows-Security-SPP"
Guid="{E23B33B0-C8C9-472C-A5F9-F2BDFEA0F156}"
EventSourceName="Software Protection Platform Service" />
    <EventID Qualifiers="16384">16384</EventID>
    <Version>0</Version>
    <Level>4</Level>
```

Moving log files to backend



- Logs are written to known location in the file system
- An agent runs on server.
- Agent sends logs periodically to backend.
- It can then clean up the log file so it doesn’t get too large.

©Len Bass and John Klein 2022

Discussion questions

1. How does the service decide to generate a log entry?
2. If time is inaccurate across computers, why put a time stamp in a log entry?

©Len Bass and John Klein 2022

Outline

- Overview
- Logs
- **Metrics**
- Tracing

Metric collection

- Metrics measure utilization of resources by VMs or containers.
- Logs are service specific. You know which service generated the log entry
- Metrics are resource specific. They can be tied to particular VM or container but not to service running inside the VM or container.
- Metrics are automatically collected by the infrastructure.
- A backend ingestion process takes the data collected by the infrastructure and transfers it to the backend database.

Metric based information

- Utilization has the form value/time period. It only makes sense over a time period.
  - Knowing that at time X, the CPU was busy does not tell you whether the CPU is overloaded. As long as it is alive, it will be busy a some point.
  - You need to know that over a time period, the CPU was busy for some percentage of the period. Similarly for other resources.
- The infrastructure will sample the VM or container to get an instantons reading and then aggregate these readings to get utilization.

©Len Bass and John Klein 2022

Discussion questions

1. Determine the information collected by AWS CloudWatch or your equivalent cloud provider and how that information can be ingested by Logstash.
2. How do you determine the correct time interval to collect utilization information?

Outline

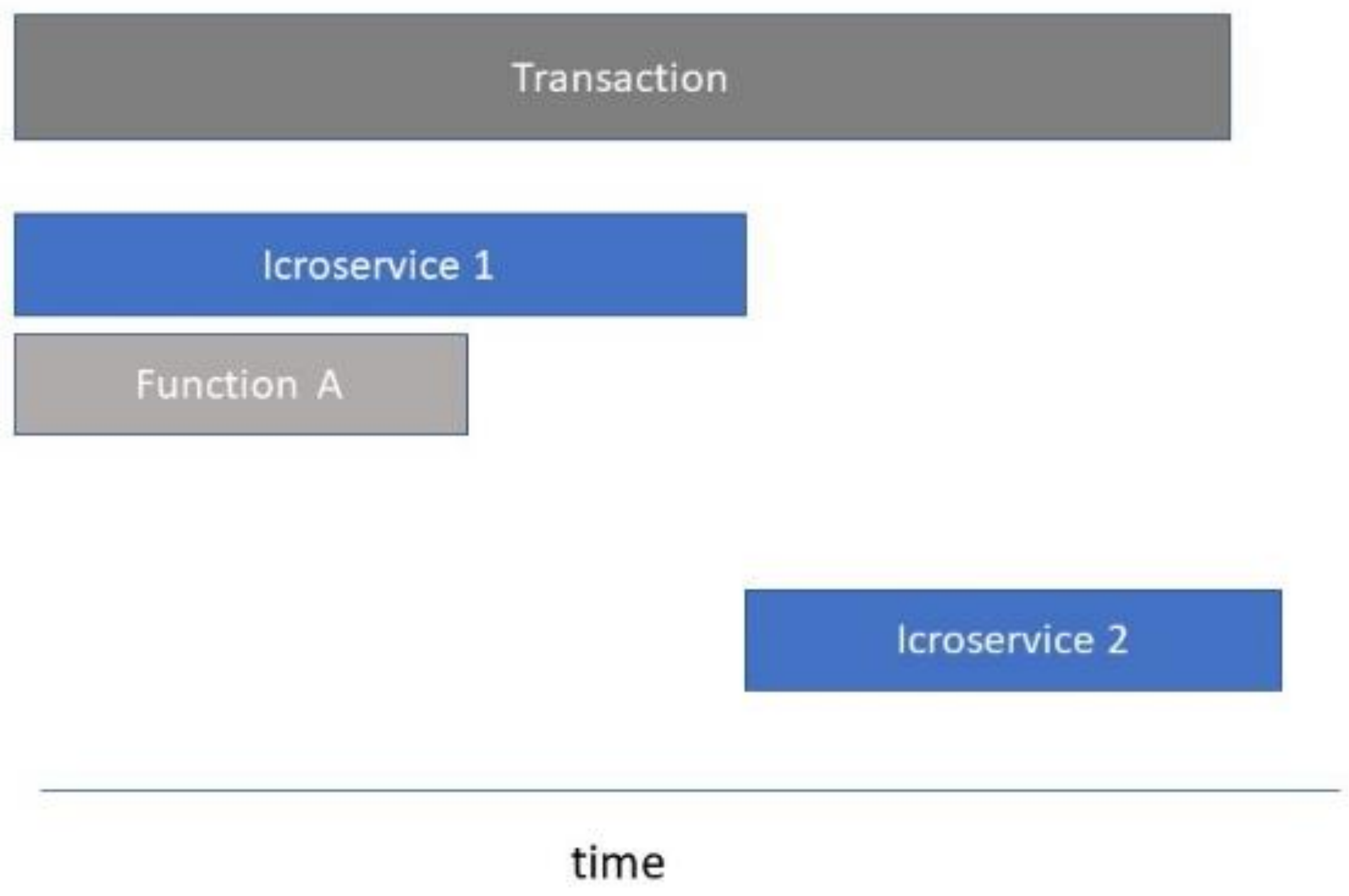
- Overview
- Logs
- Metrics
- **Tracing**

Tracing information

- Logs show single events. Traces show end to end sequences.
- A request is assigned an ID when it enters the system. The Windows log sample has an entry “Event ID”.
- This request ID is passed to each service involved in satisfying the request. This yields a request sequence.
- Request IDs are saved in the log entries of a service.
- Subsequences are called spans.

©Len Bass and John Klein 2022

Span example



Aggregating multiple requests gives information about where requests spend time. I.e. it helps to find bottlenecks.

Context

- The request ID can also be used to identify the context of a request.
- Context can be used
  - For routing
  - For A/B or canary testing
  - For system identification within a family of systems.
  - For traffic prioritization

©Len Bass and John Klein 2022

Discussion questions

1. What are the different methods available to an analyst to determine performance issues?
2. How would context be used to support canary testing?

©Len Bass and John Klein 2022