

Splunk 입문과정

2022. 4. 23

정 준 수 Ph.D

과정 목표 : Splunk 입문과정

다양한 소스의 머신 데이터(machine data)를 실시간으로 수집, 분석, 운영하는 인텔리전스 플랫폼인 Splunk의 이해를 바탕으로, 비즈니스 유형에 관계 없이 모든 장비/모든 서버/모든 장치에서 정형 · 비정형 데이터를 수집할 수 있으며, 수집된 데이터를 규칙에 따라 분석하고, 분석된 자료를 기반으로 시각화하여 사용자에게 객관적인 지표를 제공함.

1. Machine Data 의 이해
2. Splunk Architecture(구성)
3. Splunk 검색 – SPL(Splunk Processing Language)
4. Splunk 시각화(Dashboard)
5. Splunk ES 활용

강의 자료

[**https://github.com/JSJeong-me/Splunk-Monitoring**](https://github.com/JSJeong-me/Splunk-Monitoring)

**Splunk Enterprise is the data collection, indexing,
and visualization engine for operational intelligence.**

단계별 과정 내용

1 단계: Machine Data 와 IT Service Intelligence(ITSI)

**2 단계: SIEM(Security Information & Event Management)
User Behavior Analytics(UBA)
Hadoop Ecosystem**

3 단계: Machine Learning 과 Predictive Analytics

Big Data

Machine Data

Log parsing

Raw Log

Mar 19 2020 23:53:23:
%ASA-4-338002: Dynamic Filter
monitored blacklisted UDP traffic
from
Sample_Outside:238.134.165.47/
39266 (238.134.165.47/39266) to
Sample_LinuxDB:238.134.165.47/
38274

Log Parser



Parsed Log

Field names	Field values
timestamp	Mar 19 2020 23:53:23
product	ASA
log_level	4
message_id	338002
source_zone	Sample_Outside
source_ip	238.134.165.47
source_translated_ip	238.134.165.47
dest_zone	Sample_LinuxDB
dest_ip	238.134.165.47
dest_translated_ip	238.134.165.47/38274

Downstream



Splunk와 Machine Data

- 머신 데이터는 빅데이터 중에서도 가장 급증하고 복잡한 영역임
- Splunk를 통해 머신 데이터의 가치를 모든 사용자가 업무에 활용하는 한편, 분석 중심의 SIEM으로 데이터 유출 방지
사용자 및 개체 행동 분석을 통해 알려지지 않은 위협으로부터 보안을 강화
- 머신러닝을 활용한 위협 탐지의 자동화로 보다 정확한 행동 기반의 경고를 통해 검색 시간 단축 및 신속한 검토와 해결이 가능



빠르게 얻는 통찰력

빠른 검색, 강력한 쿼리 언어
(SPL) 및 보기 좋은
대시보드를 통해 문제 해결
시간 최대 90% 단축



스트레스 없는 안정적인 확장

성능에 영향을 주지
않으면서 필요에 따라 확장
가능



강력한 기능 및 머신러닝

고급 머신 러닝, 상관관계 및
알림을 통한 노이즈 감지



확장형 생태계

필요한 데이터에 상관없이
Splunk를 통해 해당 얻기

Splunk 특징

- Splunk는 데이터의 미개발 가치를 활용하여 조직 최적화와 높은 가치의 고객 경험 제공
- 고객 상황에 맞는 Splunk 제품을 통해 머신 데이터의 가치를 활용 및 운영 효율성 달성

Security

- Splunk Enterprise Security
- Splunk Phantom
- Splunk User Behavior Analytics
- Splunk Insight for Ransomware

Core

- Splunk Enterprise
- Splunk Cloud
- Splunk Light

IT Operations

- Splunk IT Service Intelligence
- Splunk Insight for AWS Cloud Monitoring
- Splunk App for Infrastructure
- VictorOps

Business Analytics

- Splunk Business Flow

IoT

- Splunk for Industrial IoT

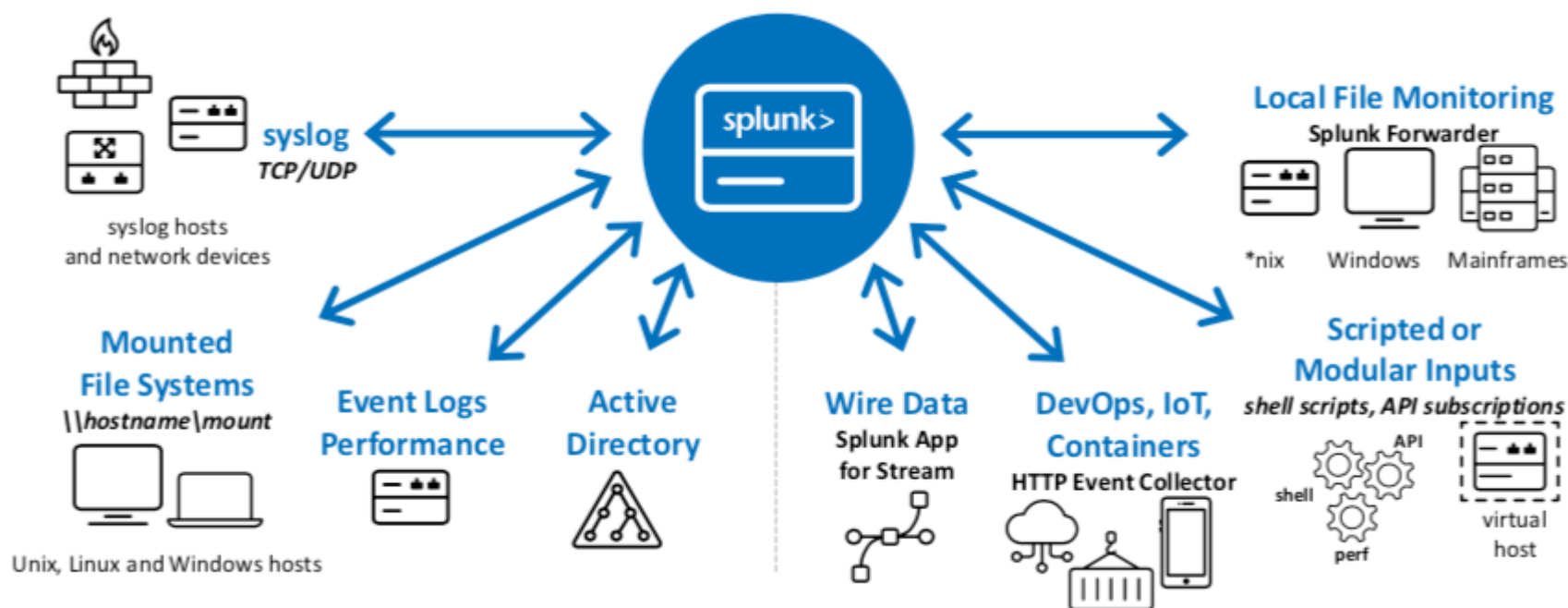
Splunk 특징: Database vs Splunk

Splunk에서 데이터베이스와 가장 유사한 개념은 인덱스(index)다. 인덱스는 비슷한 유형의 데이터 저장소이다. 인덱스에는 데이터를 저장할 수 있다. 하나의 인덱스에는 유사한 데이터들을 모으는 것이 바람직하다. 데이터베이스의 테이블과 유사한 개념은 Splunk에서는 어떤 것이 있을까? 엄밀히 말하면 테이블과 같은 구조화된 데이터 저장소는 없다고 봐야 한다.

다만 수집하는 데이터의 구조와 내용을 지정하는 sourcetype이라는 개념이 존재한다. 데이터베이스에서는 테이블이 데이터의 구조, 형식을 지정하고 테이블에 실제 데이터를 저장하였다. 하지만 Splunk는 sourcetype에 데이터를 저장하지 않는다. sourcetype은 데이터의 형식과 필드들 지정하는 일종의 파싱규칙으로 보는게 더 맞는 내용이다.

Splunk 특징

- 강력한 데이터 수집 기능 및 다양한 수집 대상: Linux/Unix/Windows 서버, 가상화 솔루션, 애플리케이션, 네트워크 장비, 보안 장비, DBMS, DRM, DLP 등
- 다양한 수집 방식: 파일시스템 모니터링, TCP/UDP 소켓, Shell Script(Shell, Python, Batch, PowerShell), Windows WMI/Registry, HTTP Event Collector Agent/Agentless 동시 지원



Splunk 특징

- 강력한 분석기능

- 140개 이상의 분석 명령어를 통한 머신 데이터에 대한 제한 없는 상관관계 분석
- 별도의 Correlation Key 설정 없이 시계열에 따른 상관관계 분석
- Lookup을 통한 외부 DB나 참조 데이터 연계 분석
- 실시간 감시 및 경고 제공: 실시간/주기적 검색 실행으로 이벤트 모니터링 및 알림 지원, 사용자 인지를 위한 RSS/E-mail/NMS/SNS 연동, 자동 대응을 위한 프로그램 및 스크립트 연동
- 머신러닝 지원: 다양한 표준 알고리즘과 라이브러리(Scikit-learn, Pandas, Scipy, Numpy, StatsModel) 적용, 모델링 지원(모델 생성, 유효성 검사, 배포)

Splunk 기대효과

- Splunk는 머신 데이터를 질문에 대한 답으로 전환하여 실시간 인사이트를 얻고, 비즈니스 성과를 향상

IT 운영

- 정교한 머신러닝으로 운영 중단을 사전에 예측 및 방지
- 신속한 데이터 기반 문제 해결로 평균 해결 시간 단축(MTTR)
- 매트릭스, 로그 및 기타 머신데이터 수집 및 상관 관계를 통해 인프라 및 어플리케이션 사전 모니터링
- 인프라 및 시스템 전반에서 데이터를 통합하여 사일로 제거
- 이상 징후 탐지 및 실시간으로 문제 예방

비즈니스 분석

- 데이터 스트림을 분석하여 패턴, 특이점 및 추세를 파악하고 End-to-End 비즈니스 프로세스에 대한 가시성 확보
- IT용 대시보드를 구축하거나 LOB 프로세스를 시각적으로 탐색하여 중요한 비즈니스 매트릭스에 대한 투명성 확보
- 매트릭스, 로그 및 기타 머신데이터 수집 및 상관 관계를 통해 인프라 및 어플리케이션 사전 모니터링
- 문제의 잠재적인 근본 원인을 조사하여 지속적인 개선 유도

보안

- 보안 및 비보안 데이터 원본에서 포괄적인 보안 분석 획득
- Kill chain 방법론을 사용한 고급 위협 탐지
- 머신러닝 기반 고급 분석을 사용하여 신속한 이상 및 위협 탐지를 수행하고 내·외부자 위협을 완화
- 자동화 및 인적지원을 통해 운영 효율성을 향상시키는 적응형 대응과 패턴 플레이북

대용량 데이터 분석

- Splunk 소프트웨어를 쉽게 배포하고 사용
- 빅데이터 요구에 맞게 솔루션 확장
- 실시간 및 이전 데이터 검색
- 하나의 통합 뷰에서 다양한 데이터세트 분석

Splunk Product

Splunk Enterprise

실시간 가시성

중요한 머신 데이터의 수집,
인덱싱 및 경고 자동화

데이터 소스의 무제한

소스 또는 형식에 상관없이 모든
데이터에서 실행 가능한 통찰력
발견

AI 및 머신러닝

예측 및 사전대응의 비즈니스
의사 결정을 위한 AI와
머신러닝의 활용

Splunk Enterprise Security

탐지 시간 단축

클라우드와 온-프레미스 환경에
상관없이, 머신데이터를
파악하여 사용자 환경의 위협을
신속하게 탐지하는 완벽한
가시성을 제공

조사 간소화

하나의 중앙집중식 뷰에서
잠재적으로 보안에 관련된
사항들을 조사

신속한 대응

자동화 된 작업 및 워크 플로우를
사용하여 신속하고 적절하게
대응

Splunk User Behavior Analytics

지능적 위협 탐지

기존의 보안툴이 놓치는 이상 및
알려지지 않은 위협을 발견

생산성 향상

수백 가지 이상 징후를 하나의
위협으로 자동화하여 처리

위협 검색의 가속화

모든 개체의 이상 또는 위협에
대한 심층 조사 기능 및 강력한
행동 베이스라인 사용

Splunk IT Service Intelligence

예측 및 방지

영향이 발생하기 전에 팀이
신속하게 움직일 수 있도록
함으로써 서비스 저하 예측 및
조사 진행

확장 가능한 데이터 플랫폼

대량의 정형 및 비정형 데이터를
실시간으로 처리

360도 인사이트

AIOps 기반 통찰력을 활용하여
문제 탐지, 조사 간소화, 문제
분류 및 해결 가속화

Splunk Components

Splunk is comprised of three main processing components:



Indexer

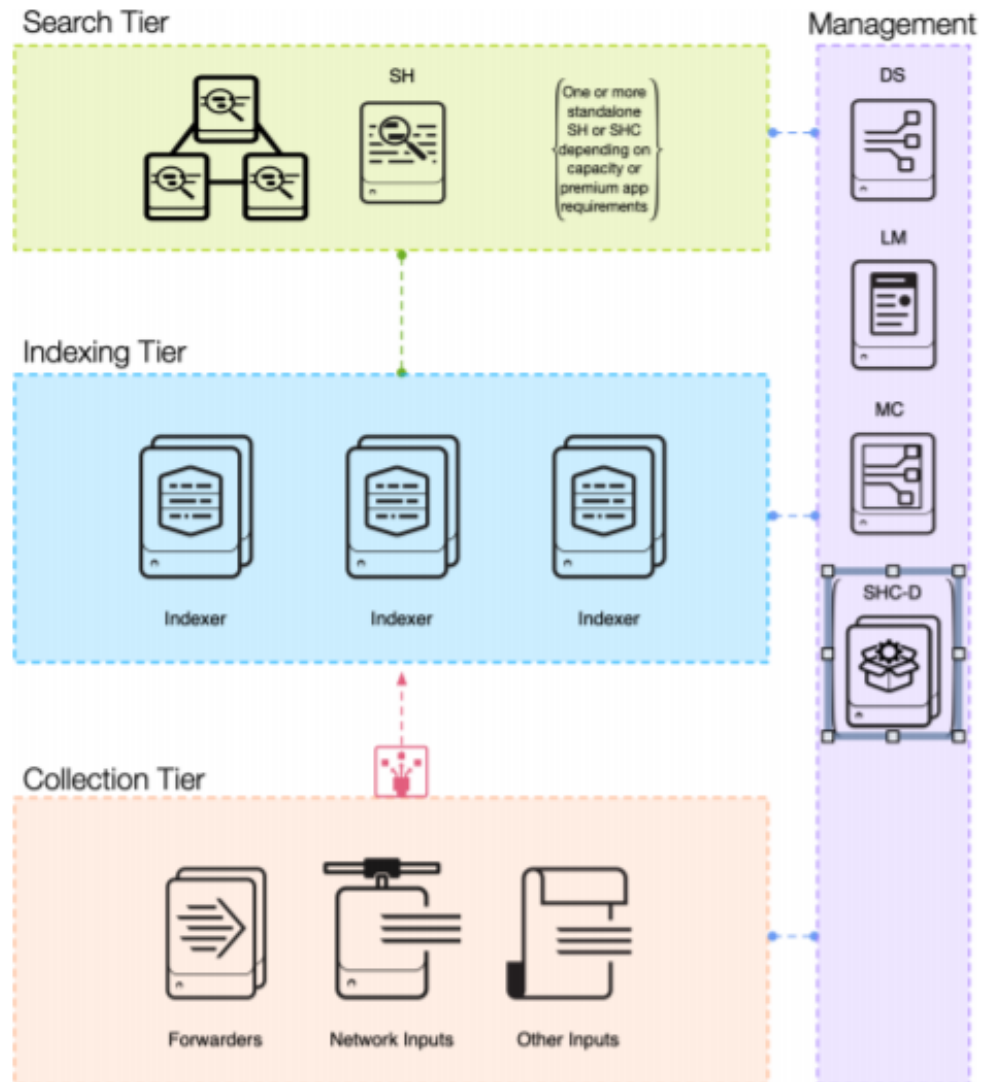


Search Head

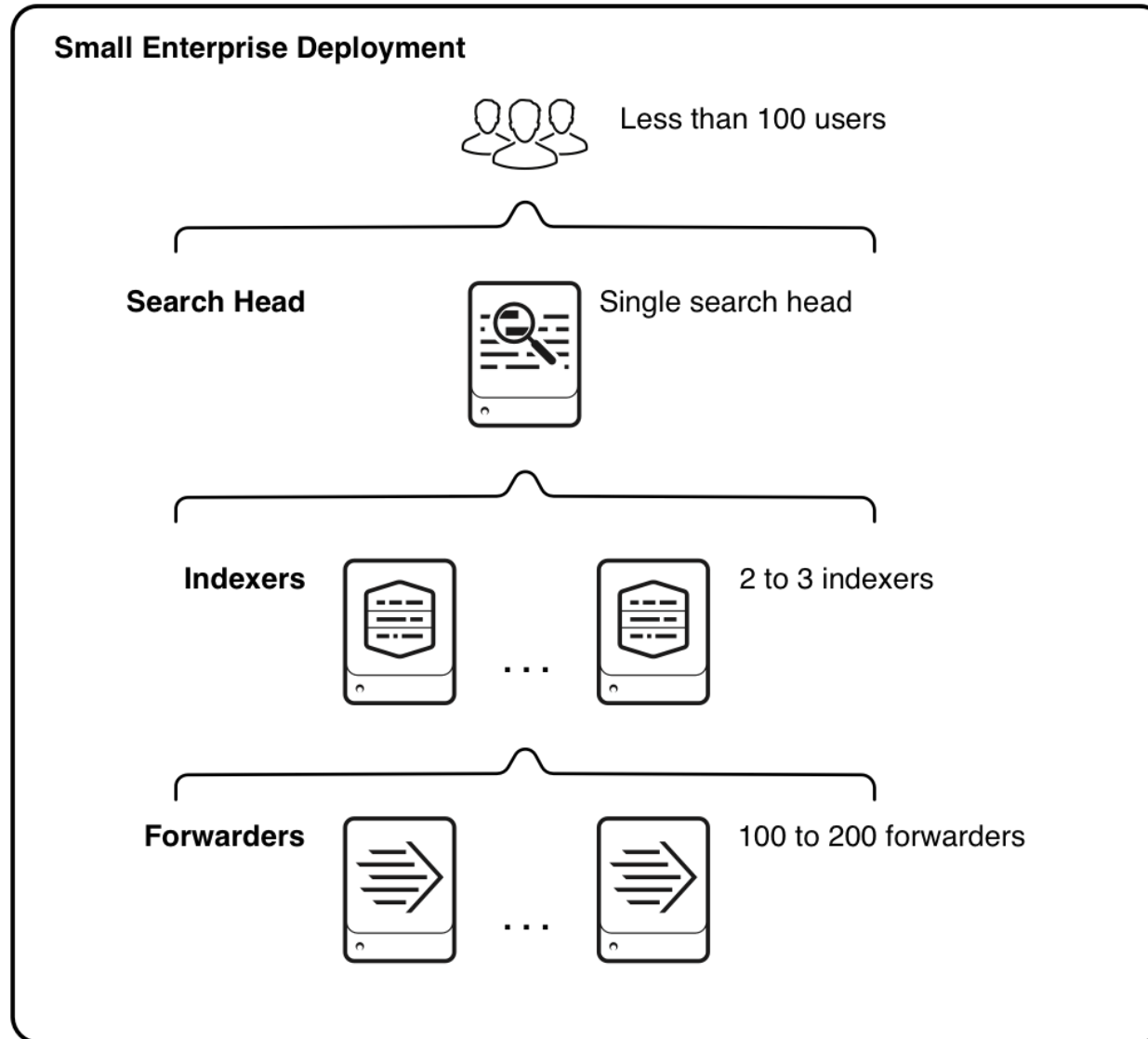


Forwarder

Splunk Architecture



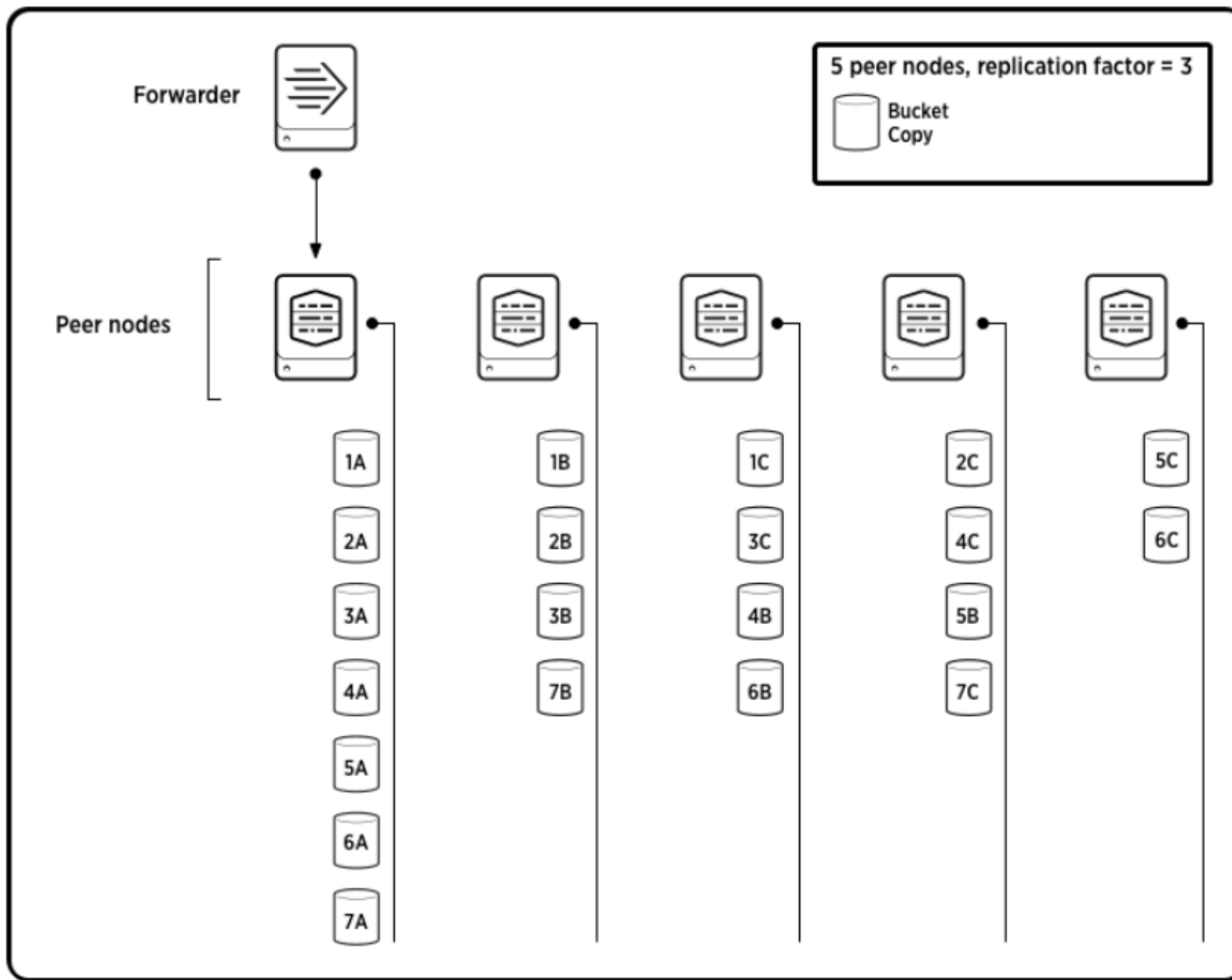
Small enterprise deployment



Splunk Indexing

What data can I index?

The process of transforming the data is called **indexing**. During indexing, the incoming data is processed to enable fast searching and analysis. The processed results are stored in the index as **events**.



Data Collection Components

(UF) Universal Forwarder

The UF provides:

- Checkpoint/restart function for lossless data collection
- Efficient protocol that minimizes network bandwidth utilization
- Throttling capabilities
- Built-in, load-balancing across available indexers
- Optional network encryption using SSL/TLS
- Data compression (use only without SSL/TLS)
- Multiple input methods (files, Windows Event logs, network inputs, scripted inputs)
- Limited event filtering capabilities (Windows event logs only)
- Parallel ingestion pipeline support to increase throughput/reduce latency

Data Collection Components

(HF) Heavy Forwarder

The HF:

- Parses data into events
- Filters and routes based on individual event data
- Has a larger resource footprint than the UF
- Has a larger network bandwidth footprint than the UF (up to 5x)
- Provides a GUI for management

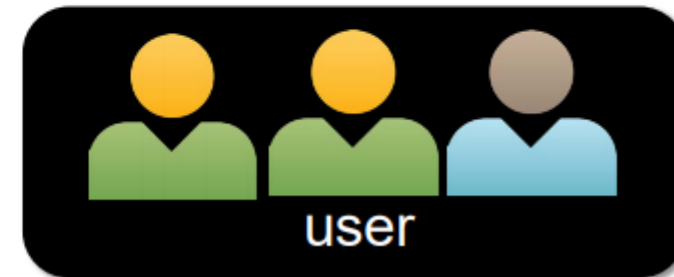
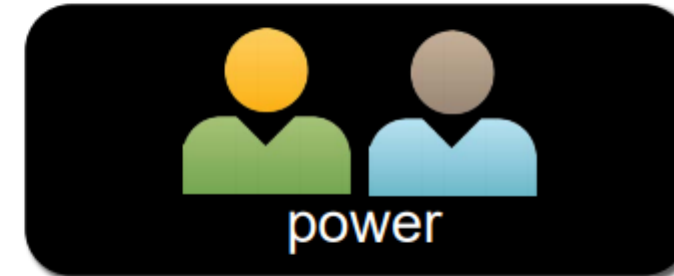
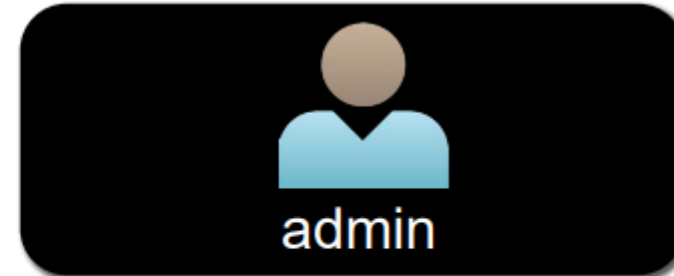
Users and Roles

- Splunk users are assigned roles, which determine their capabilities and data access
- Out of the box, there are 3 main roles:
 - Admin
 - Power
 - User
- Splunk admins can create additional roles

Note

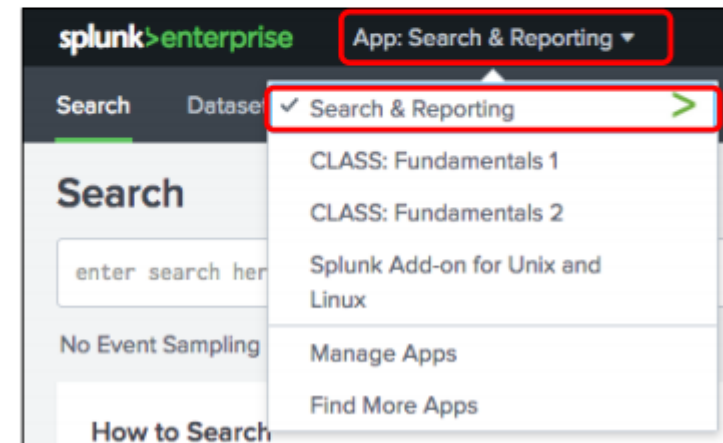
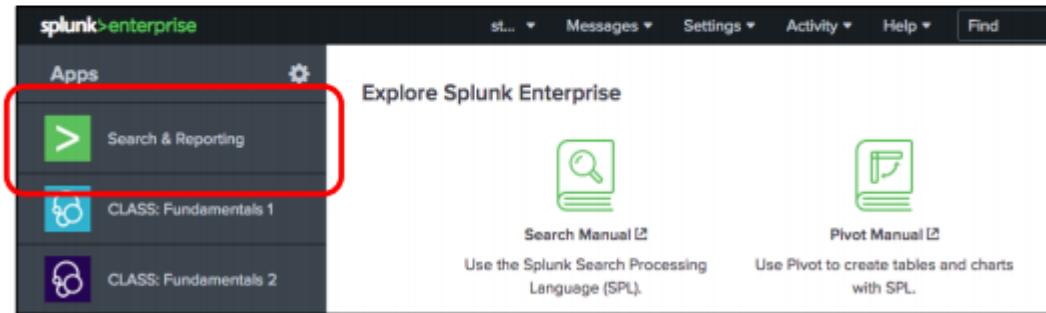


In this class, the account you'll use for the lab exercises has been assigned the **Power** role.

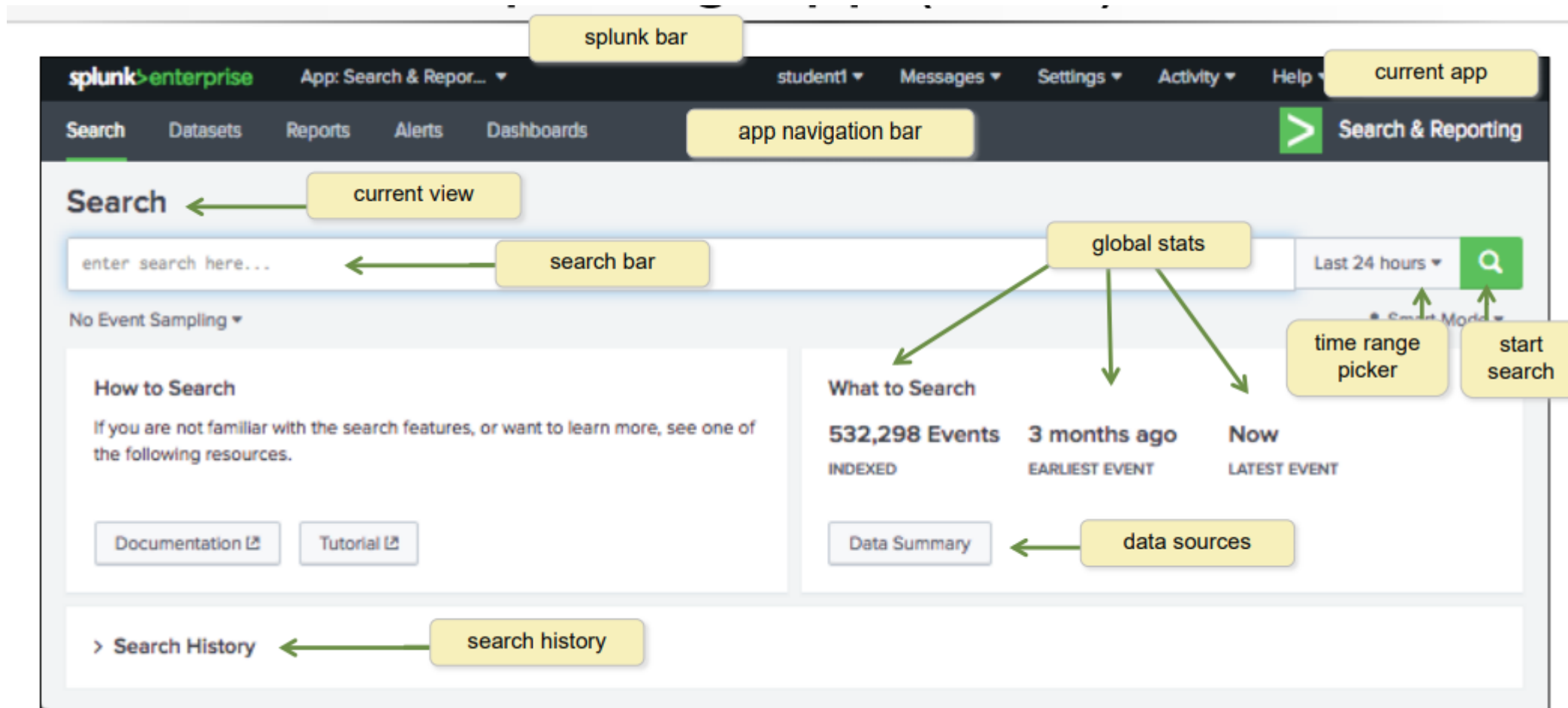


Search & Reporting App

- Provides a default interface for searching and analyzing data
- Enables you to create knowledge objects, reports, and dashboards
- Access by selecting the **Search & Reporting** button on the Home app or from an app view, select **Apps > Search & Reporting**



Search & Reporting App



Data Summary Tabs

The screenshot shows the Splunk Enterprise interface. The top navigation bar includes 'Search', 'Datasets', 'Reports', 'Alerts', and 'Dashboards'. The 'Search' section has a search bar and a 'Last 24 hours' filter. The 'What to Search' section shows '532,486 Events' and '3 months ago'. A green box highlights the 'Data Summary' link. Three overlapping panels show the 'Data Summary' tabs for 'Hosts (10)', 'Sources (15)', and 'Sourcetypes (10)'. The 'Sourcetypes' panel is the most prominent, showing a table with columns for 'Sourcetype', 'Count', and 'Last Update'. A yellow callout box points to the 'Data Summary' link with the text 'Click Data Summary to see hosts, sources, or sourcetypes on separate tabs'. Another yellow callout box points to the 'Sourcetypes' table with the text 'Tables can be sorted or filtered'.

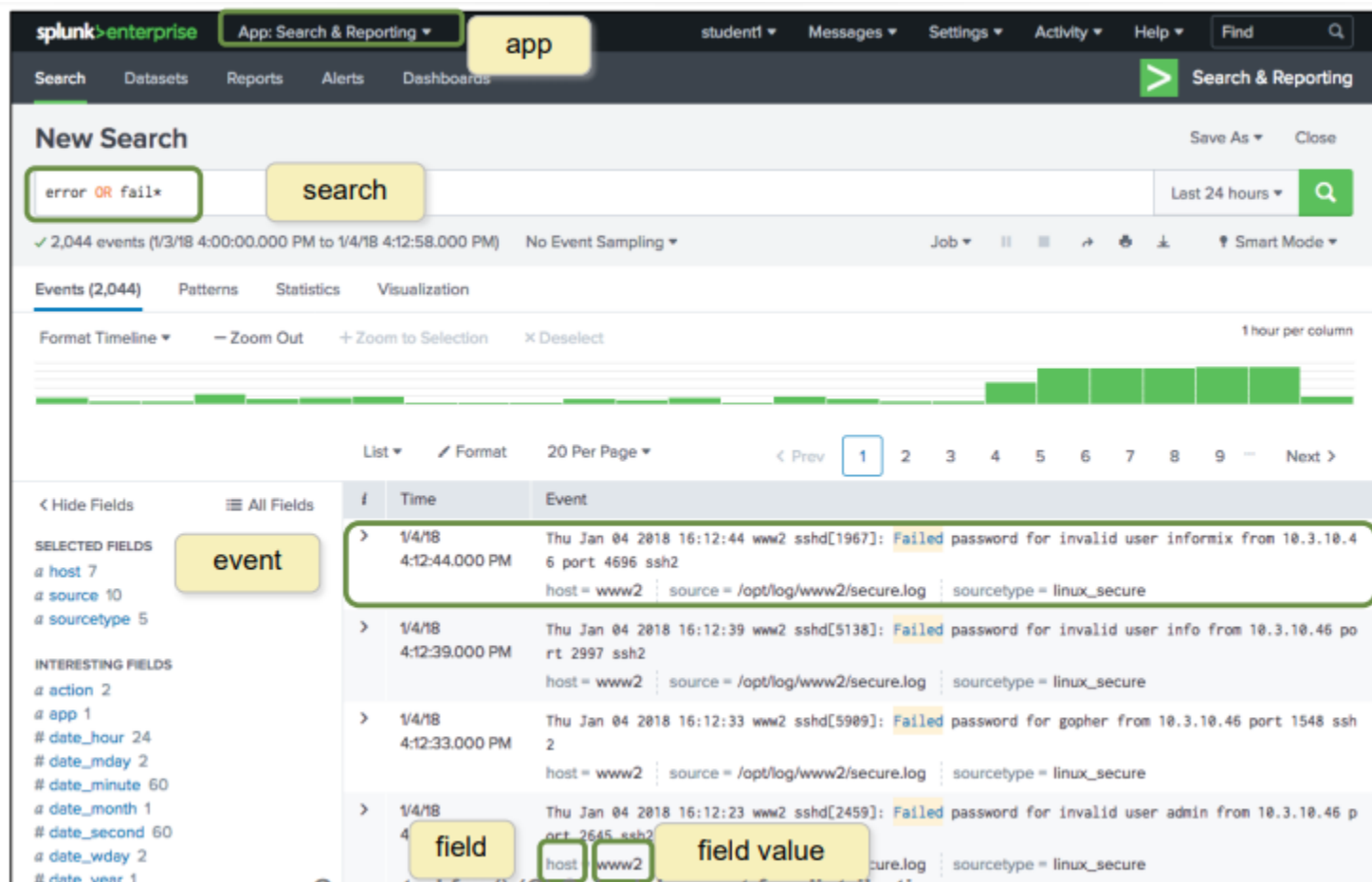
Click **Data Summary** to see hosts, sources, or sourcetypes on separate tabs

- Host – Unique identifier of where the events originated (host name, IP address, etc.)
- Source - Name of the file, stream, or other input
- Sourcetype - Specific data type or data format

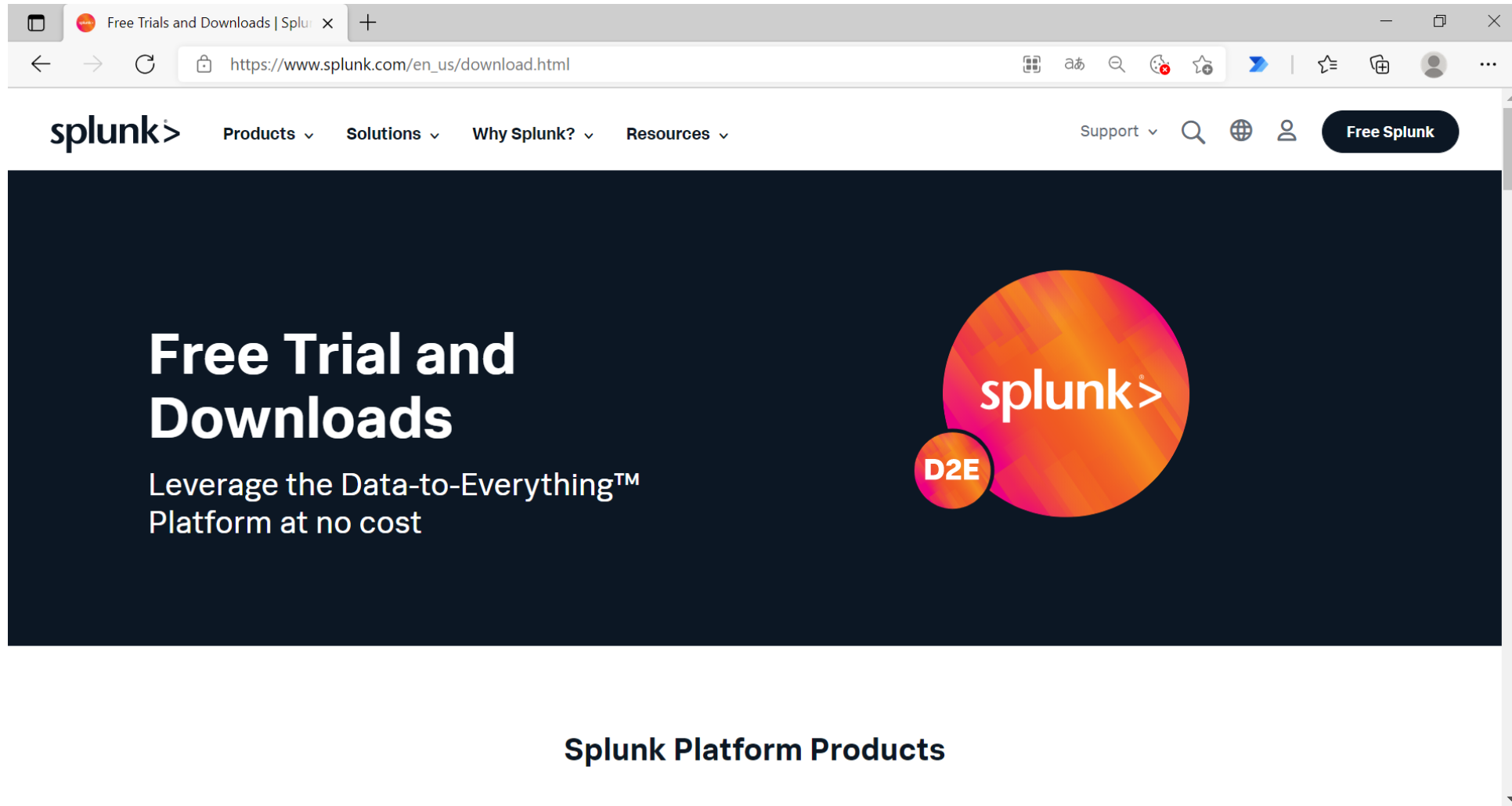
Tables can be sorted or filtered

Sourcetype	Count	Last Update
SimCubeBeta	377	1/4/18 3:51:45.000 PM
access_combined	154,373	1/4/18 3:52:18.000 PM
cisco_esa	3,200	1/4/18 3:52:15.000 PM
cisco_firewall	538	1/4/18 10:23:47.000 AM
cisco_wsa_squid	3,749	1/4/18 3:50:37.000 PM
history_access	7,662	1/4/18 10:23:46.000 AM
linux_secure	16,950	1/4/18 3:52:12.000 PM
sales_entries	215,869	1/4/18 3:51:56.000 PM
vendor_sales	120,459	1/4/18 3:49:32.000 PM
winauthentication_security	9,372	1/4/18 10:23:46.000 AM

Events Tab

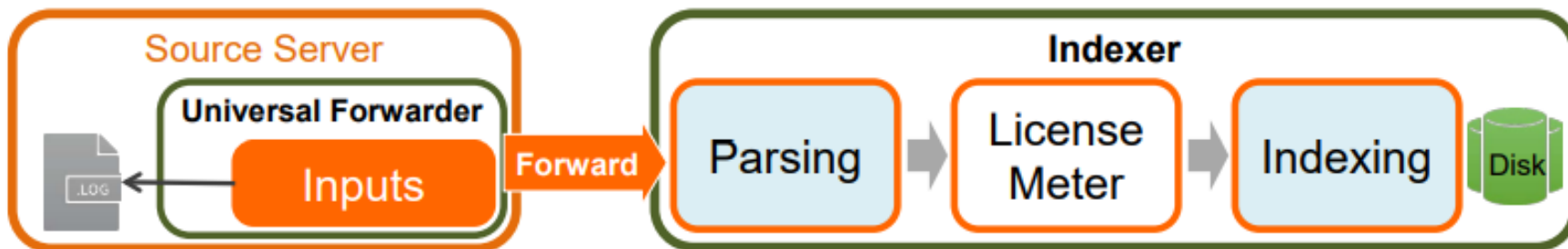


Splunk Installation



Splunk Index 생성과정

- Splunk index time process (data ingestion) can be broken down into three phases:
 1. **Input phase:** handled at the source (usually a forwarder)
 - The data sources are being opened and read
 - Data is handled as streams and any configuration settings are applied to the entire stream
 2. **Parsing phase:** handled by indexers (or heavy forwarders)
 - Data is broken up into events and advanced processing can be performed
 3. **Indexing phase:**
 - License meter runs as data and is initially written to disk, prior to compression
 - After data is written to disk, it **cannot** be changed

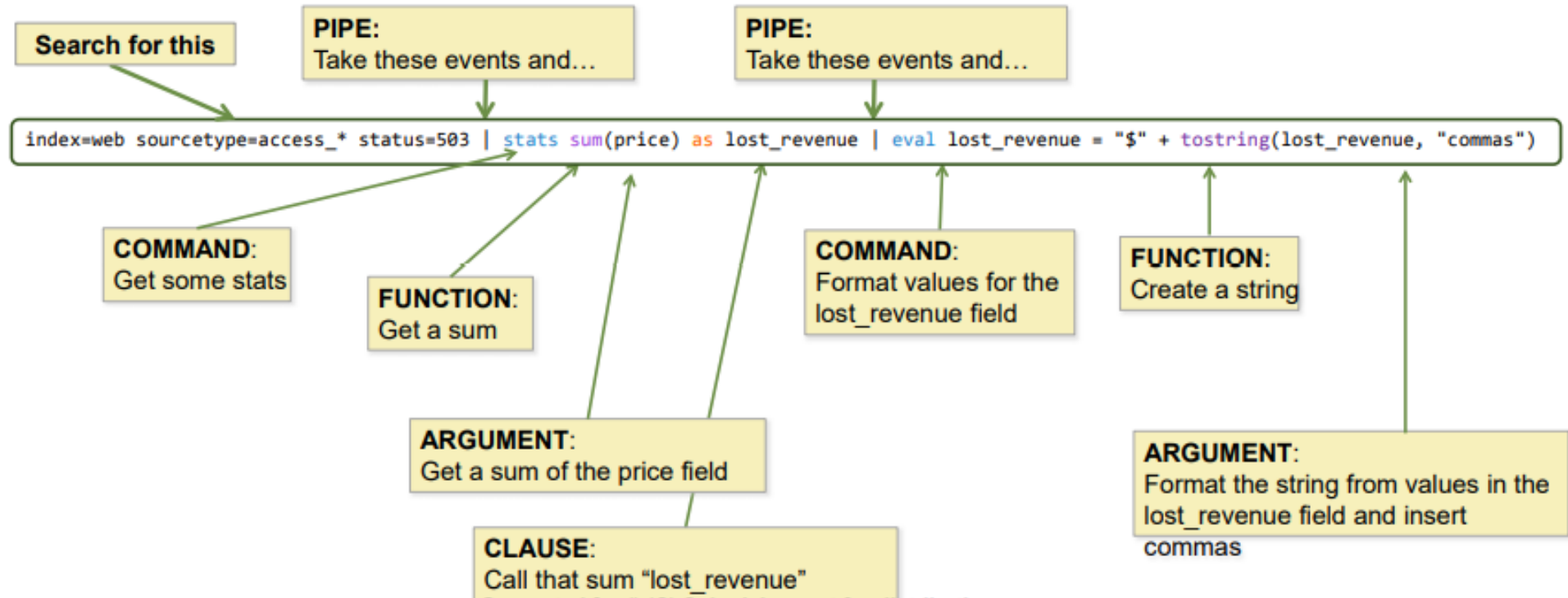


Data Input Types

- Splunk supports many types of data input
 - **Files and directories:** monitoring text files and/or directory structures containing text files
 - **Network data:** listening on a port for network data
 - **Script output:** executing a script and using the output from the script as the input
 - **Windows logs:** monitoring Windows event logs, Active Directory, etc.
 - **HTTP:** using the HTTP Event Collector
 - And more...
- You can add data inputs with:
 - Apps and add-ons from Splunkbase
 - Splunk Web
 - CLI
 - Directly editing `inputs.conf`

Splunk Language Syntax

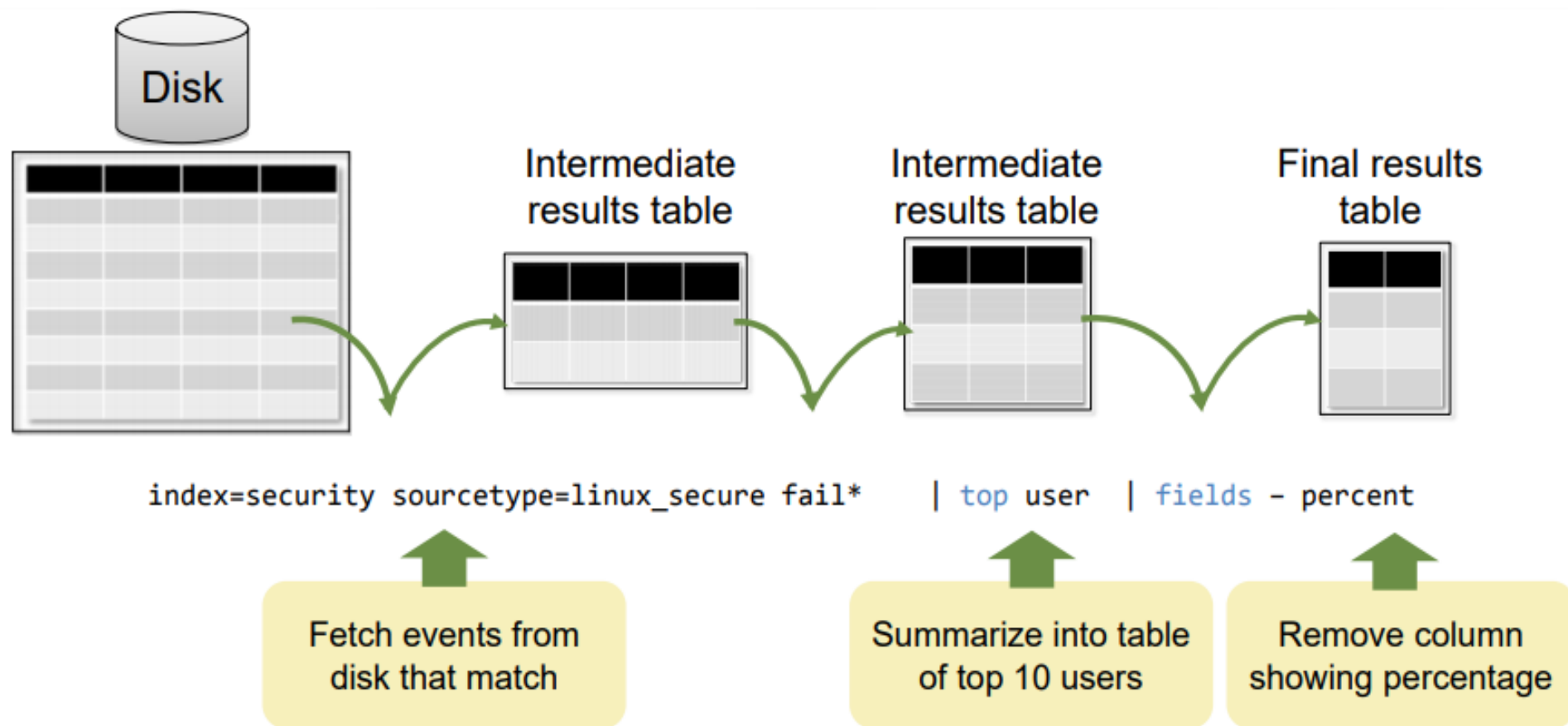
This diagram represents a search, broken into its syntax components:



Splunk Language Syntax Components

- Searches are made up of 5 basic components
 1. **Search terms** – what are you looking for?
 - Keywords, phrases, Booleans, etc.
 2. **Commands** – what do you want to do with the results?
 - Create a chart, compute statistics, evaluate and format, etc.
 3. **Functions** – how do you want to chart, compute, or evaluate the results?
 - Get a sum, get an average, transform the values, etc.
 4. **Arguments** – are there variables you want to apply to this function?
 - Calculate average value for a specific field, convert milliseconds to seconds, etc.
 5. **Clauses** – how do you want to group or rename the fields in the results?
 - Give a field another name or group values by or over

The Search Pipeline



Creating a Report

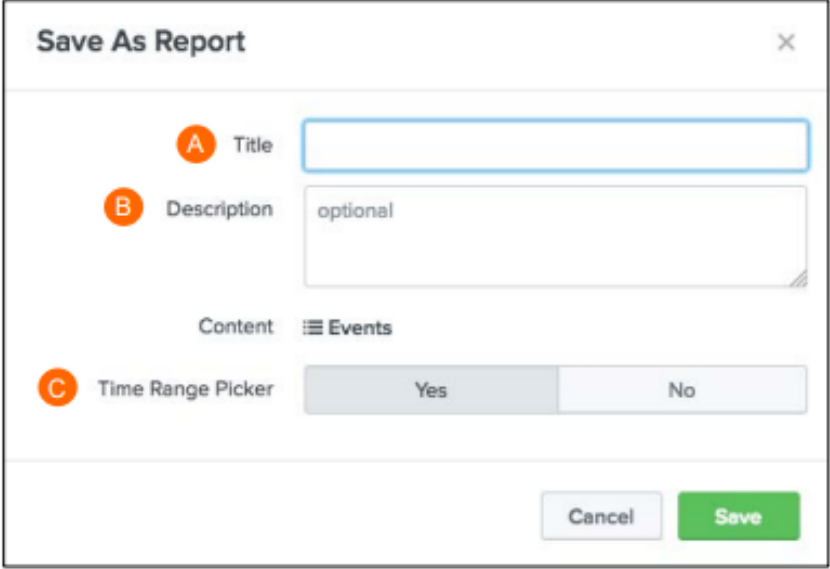
- 1 Run a search
- 2 Select **Save As**
- 3 Select **Report**

The screenshot shows the Splunk Search interface. At the top, the search bar contains the query: `index=web sourcetype=access_combined action=purchase status!=200`. Below the search bar, a status bar indicates "50 events (3/11/18 7:00:00.000 PM to 3/12/18 7:06:05.000 PM)". The main area features a timeline visualization with green bars representing event counts over time. Below the timeline, there is a table of search results. The table has columns for "Time" and "Event". The first row shows a timestamp of "3/12/18 6:09:24.000 PM" and an event description starting with "81.11.191.113 - - [12/Mar/2018:18:09:24] \"POST /cart.do?action=purchase&itemId=EST-17&JSESSIONID=SD10SL9FF5ADFF4963 HTTP 1.1\" 503 2768 \"http://www.buttercupgames.com/cart.do?action=addtocart&itemId=EST-17&categoryId=ARCADE&productId=MB-AG-G07\" \"Googlebot/2.1 (http://www.googlebot.com/bot.html)\" 846".

Time	Event
3/12/18 6:09:24.000 PM	81.11.191.113 - - [12/Mar/2018:18:09:24] "POST /cart.do?action=purchase&itemId=EST-17&JSESSIONID=SD10SL9FF5ADFF4963 HTTP 1.1" 503 2768 "http://www.buttercupgames.com/cart.do?action=addtocart&itemId=EST-17&categoryId=ARCADE&productId=MB-AG-G07" "Googlebot/2.1 (http://www.googlebot.com/bot.html)" 846

Creating a Report

- A Give the report a meaningful title (required)
- B Specify a description (optional)
- C Select whether to include or not to include a time range picker
 - The report is saved with the time range that was selected when it was created
 - Adding a time range picker allows you to adjust the time range of the report when you run it



The image shows a 'Save As Report' dialog box with a close button (X) in the top right corner. It contains three main sections: 'Title' with a text input field, 'Description' with a text area containing the word 'optional', and 'Content' with a dropdown menu set to 'Events'. Below these is a 'Time Range Picker' section with two radio buttons, 'Yes' and 'No', where 'Yes' is selected. At the bottom right are 'Cancel' and 'Save' buttons.

Save As Report

A Title

B Description

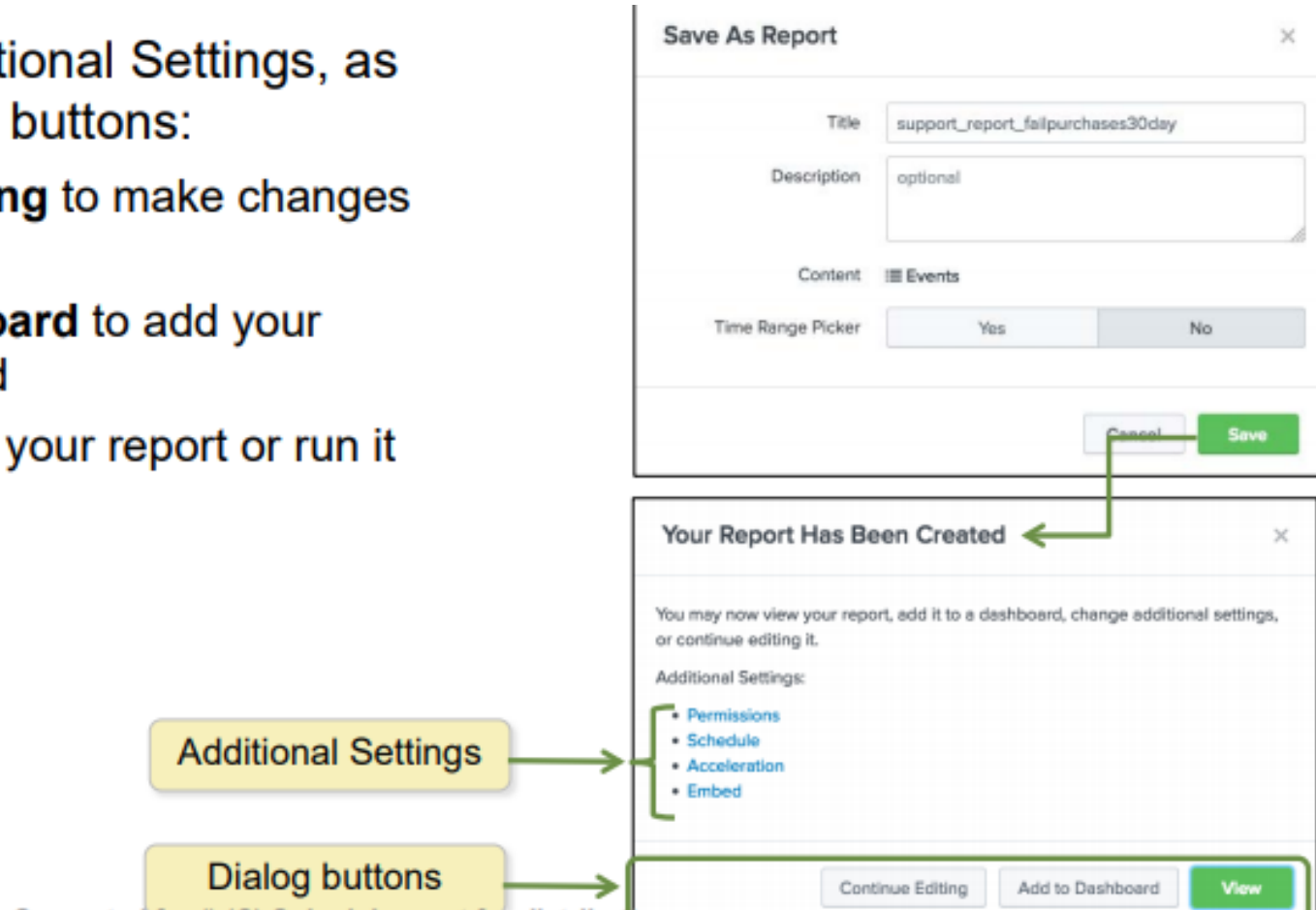
Content

C Time Range Picker ☒ Yes ☐ No

Cancel Save

Creating a Report

- You can change Additional Settings, as well as use the dialog buttons:
 - Click **Continue Editing** to make changes to your report
 - Click **Add to Dashboard** to add your report to a dashboard
 - Click **View** to display your report or run it again



기본 search

error OR fail*

error 이 포함된것과 fail*이 포함된것을 검색한다.

기본 search - AND

fail* password

공백에는 AND가 포함된것이다.

time

s=seconds, m=minus, h=hours, d=days, w=week, mon = months, y=year

시간은 keyword로 대표된다.

fast(고속) & smart & verbose(상세모드) 차이

field discover 진행의 유무로 구분된다.

field discover: 해당 field의 통계값

!= 와 NOT 에는 차이가 있다.

우선순위

OR가 가장 높아요

나머지는 오는 순서대로입니당.

$a=1 \text{ OR } a=2 \text{ } b=3 \text{ } (a=1 \text{ OR } a=2) \text{ } b=3$

Password fail* | stats count by eventtype

Tag This Field

Tags -> ok

Tag this field

Field -> eventtype=risky

Tags -> not_ok

Searching

Tag=not_ok

정 준 수 / Ph.D (jsjeong@hansung.ac.kr)

- 前) 삼성전자 연구원
- 前) 삼성의료원 (삼성생명과학연구소)
- 前) 삼성SDS (정보기술연구소)
- 現) (사)한국인공지능협회, AI, 머신러닝 강의
- 現) 한국소프트웨어산업협회, AI, 머신러닝 강의
- 現) 서울디지털재단, AI 자문위원
- 現) 한성대학교 교수(겸)
- 전문분야: Computer Vision, 머신러닝(ML), RPA
- <https://github.com/JSJeong-me/>

