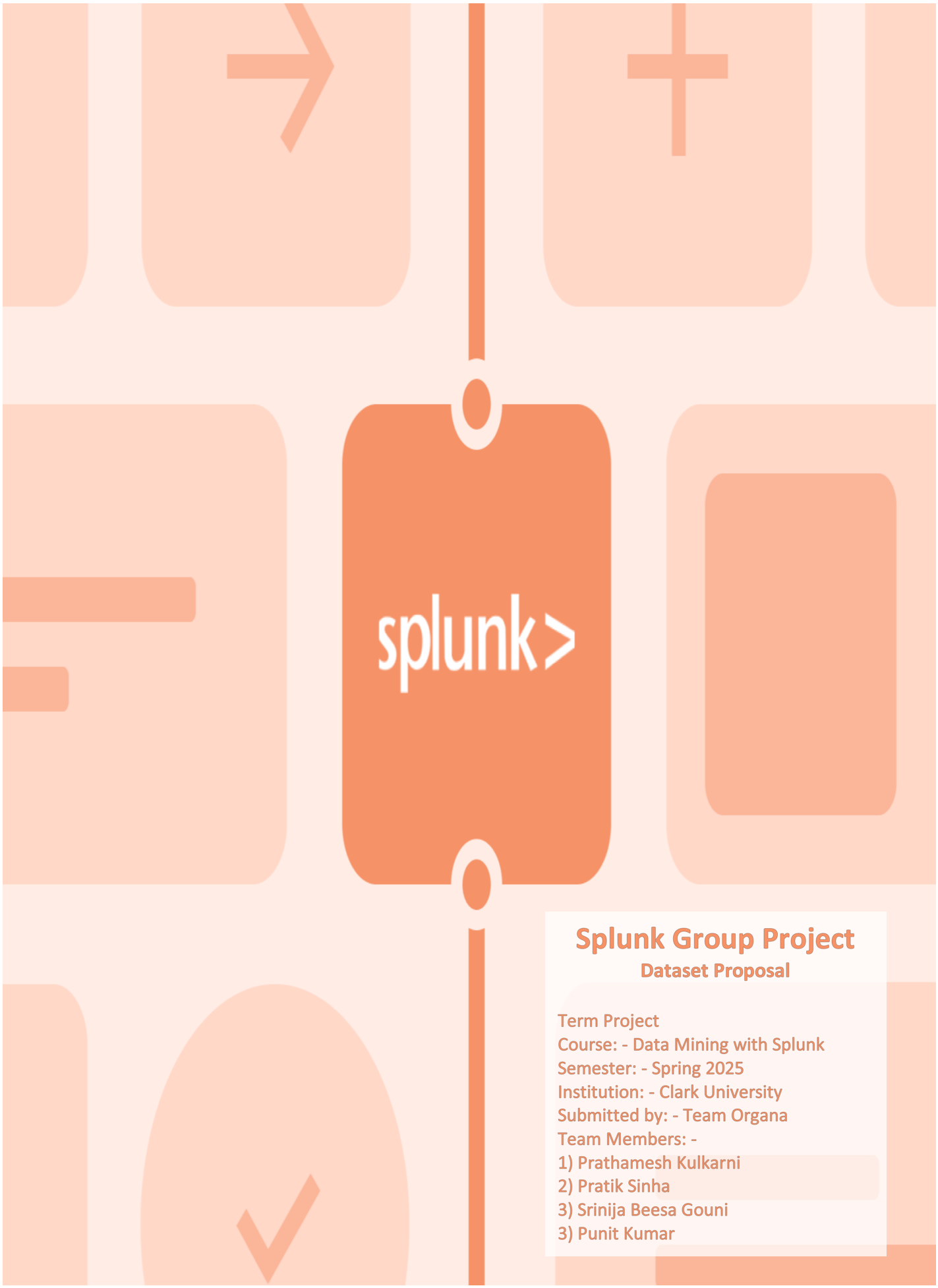

Splunk Group Project

Dataset Proposal

Term Project
Course: - Data Mining with Splunk
Semester: - Spring 2025
Institution: - Clark University
Submitted by: - Team Organa
Team Members: -

1) Prathamesh Kulkarni
2) Pratik Sinha
3) Srinija Beesa Gouni
3) Punit Kumar

Dataset 1:

Source: ToN_IoT Datasets (Telemetry Logs)

Download Link: https://research.unsw.edu.au/projects/toniot-datasets

Size/Number of Records:
- Approximately 600+ MB of data
- Millions of event records capturing IoT traffic data.
- We will only use the IoT traffic for our analysis; and exclude network and OS logs.

5Vs Analysis:
- Volume: High; IoT log size exceeds 600 MB.
- Velocity: High; data collected continuously, simulating real-time event frequencies.
- Variety: Semi-structured and unstructured; includes log files.
- Veracity: Very high; generated by credible cybersecurity research at UNSW Canberra.
- Value: Very high; suitable for advanced cybersecurity analysis, threat detection, anomaly detection, and performance monitoring using Splunk.

Dataset 2:

Source: Traffic and Log Data Captured During a Cyber Defense Exercise

Download Link: https://zenodo.org/record/3746129

Size/Number of Records:
- Approximately 274.5 MB total
- Multiple files capturing two days of network traffic and event logs

5Vs Analysis:
- Volume: Medium to high; comprehensive logs over two days of exercises.
- Velocity: Moderate; logs and network flows recorded at high frequency during active exercise sessions.
- Variety: Semi-structured; includes JSON (IPFIX traffic flows) and XML (Windows Event Logs).
- Veracity: High; sourced from controlled cyber defense exercise hosted by Masaryk University.
- Value: High; valuable for analyzing cybersecurity defense mechanisms, intrusion detection systems, and network monitoring capabilities in Splunk.

Dataset 3:

Source: Detecting Malicious URLs Dataset

Download Link: https://www.sysnet.ucsd.edu/projects/url/

Size/Number of Records:
- Approximately 470 MB (Matlab format)
- Contains about 2.4 million URLs.

5Vs Analysis:
- Volume: High; substantial dataset with millions of URLs and features.
- Velocity: Low; static dataset capturing URLs over a specific period.
- Variety: Semi-structured; available in Matlab format.
- Veracity: High; curated by researchers from the University of California, San Diego.
- Value: High; essential for developing and testing machine learning models for malicious URL detection in cybersecurity applications.

Dataset 4:

Source: ESPset: Vibration-based Fault Diagnosis of Electric Submersible Pumps

Download Link: https://data.mendeley.com/datasets/m268jsw339/1

Size/Number of Records:
- Approximately 1.3 GB
- Contains 6,032 vibration signal records with 12,103 features each.

5Vs Analysis:
- Volume: High; large dataset with detailed vibration signal records.
- Velocity: Moderate; data collected during specific test conditions, not in real-time.
- Variety: Structured; data available in CSV format with numerical features. However, this is a real life dataset, with predictive maintenance application in the asset-intensive industries. That justifies the inclusion of this dataset in this proposal.
- Veracity: High; collected from real-world ESPs used in offshore oil exploration.
- Value: High; valuable for developing predictive maintenance models and analyzing equipment health in industrial settings using Splunk.