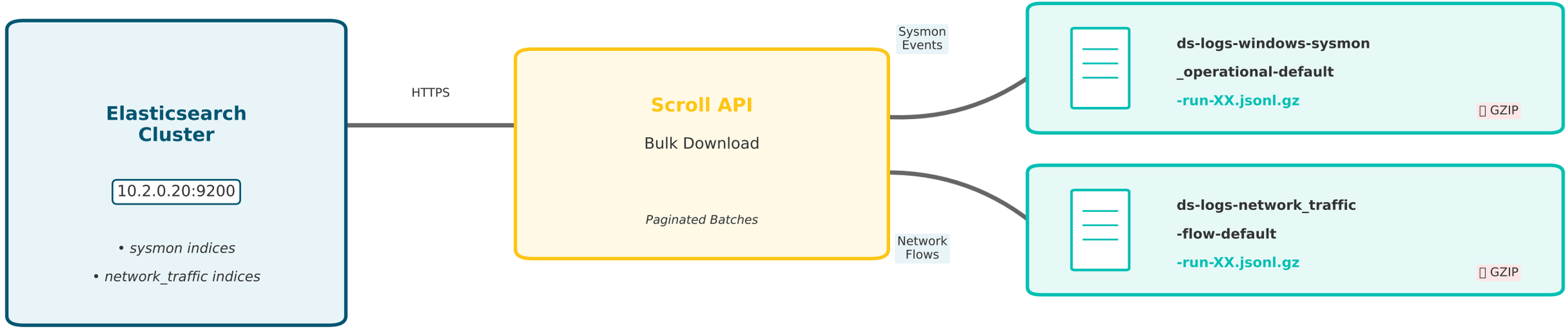
Data Source
Processing
Output Files

Figure 1.1: Elasticsearch Data Extraction Architecture

Pipeline Step 1: Raw Data Extraction from Monitoring Infrastructure



Typical Dataset Characteristics

• Volume: 100K-1M+ events per APT run • Compression: ~60-80% size reduction (GZIP) • Format: JSON Lines (JSONL)
• Sources: Windows Sysmon (host events) + Network Traffic Flow (network events) • Authentication: Username/password • SSL: Disabled (lab environment)