Cygnus

Brennon York

What is Cygnus?

- Compiler
  - Input
    - .conf with parse definition
  - Outputs
    - .java
    - .java & .class (compiled)
    - .jar (self executing)
- Generates Java Ingestors based on user-defined configuration file

Why?

- Abstracts ingest process to what is needed
  - Schema definitions and data structures
- Removes confusion of Accumulo implementation details
  - Maintains Accumulo functionality and features (sharding, visibility, etc.)
- Quickly get the data desired into Accumulo

Caveats

- Performance
  - Further testing against current Ingest speeds
  - Currently looking at ~20-30% overhead from a “fully” optimized Ingest
- ???

Features

- Streaming
  - Every java output has a “-s” or “--streaming” option to stream input from stdin
- Protobufs
  - Option to output a serialized object of the schema for iterators and front-end capability
- Threading
  - User-defined thread count value

Features

- Maximum Records
  - Define a maximum record size (per thread) to read in before writing out the Rfiles
    - Smaller number for better performance on Ingest
    - Larger number for smaller number of Accumulo major compactions
- Column Visibility
  - User-defined visibility for individual schemas

Features

- Timestamp
  - User-defined timestamp field for age-off functionality
- String-delimited Parsing
  - Allows for greater flexibility of ascii-based data structures
- Dynamic Record Sizing
  - Read previously defined value as a size

Parse Definition

```
%8[ts]%4[dur] # Time-based fields
```

- % is key
  - The % defines a new Const – Label pair
- Constants, Labels, and 'S'
  - Static values (i.e. Constants such as 1, 3, 45, etc.)
  - Previously-defined values as constants (Labels)
  - Arbitrary string ('S')
    - String delimitation (not char)

```
%S" "[date]%S" "[time]
```

- Labels
  - Brackets house the variable label (i.e. [sip])

Schema

```
R:sip CF:dip CV:"USER_A" TS:ts VAL:appl
```

- Locality fields
  - R, CF, CQ, CV, VAL
    - Only "R" is required
- ":"
  - Defines the label set for a locality field
- "+"
  - Allows for multiple labels to be concatenated together (i.e. R:sip+dip+dport)

Sample SiLK .conf

```
TABLE_NAME=flowSchema # Table name to give within Cloudbase
HEADER_LEN = 52 # For SiLK we skip 52 bytes
THREAD_NUM = 4
LOCAL_INPUT_DIR=/home/brennon.york/cygnus_svn/java/in_dir
LOCAL_OUTPUT_DIR=/home/brennon.york/cygnus_svn/java/out_dir

%8[ts]%4[dur]             # Time-based fields
%2[sport]%2[dport]        # Ports
%1[proto]                 # Protocol
%1[ct]                    # class / flow type
%2[sensor]                # Senser
%1[flags]%1[init]%1[sess] # Flags
%1[attr]                  # Attributes
%2[appl]                  # Application
%2                        # Unused
%2[in]%2[out]             # SNMP In and Out paths
%4[packets]%4[bytes]      # Flow details
%4[sip]%4[dip]%4[nhip]    # IPs

R:sip CF:dip+dport+proto VAL:bytes+packets+"1"+dur
R:dip CF:sip+dport+proto VAL:bytes+packets+"1"+dur
```

Demonstration

Future Features

- Automatic load into Accumulo
  - Variables to allow automatic loading into Accumulo
- Sharding
  - Round robin value to append onto Row ID
- Direct HDFS Support
  - Read / write directly to / from HDFS

Future Features

- Format Support
  - Augment user-defined values with formatters for unified data types
    - Time stamps
    - IP's
- Metadata Augmentation
  - GeoIP
  - Entropy
  - Tagging (Blacklists, Whitelists, CDN's, etc.)