Развитие проекта

ATLAS EventIndex

Семинар Лаборатории Ядерных Проблем

16.05.2019

Event Index

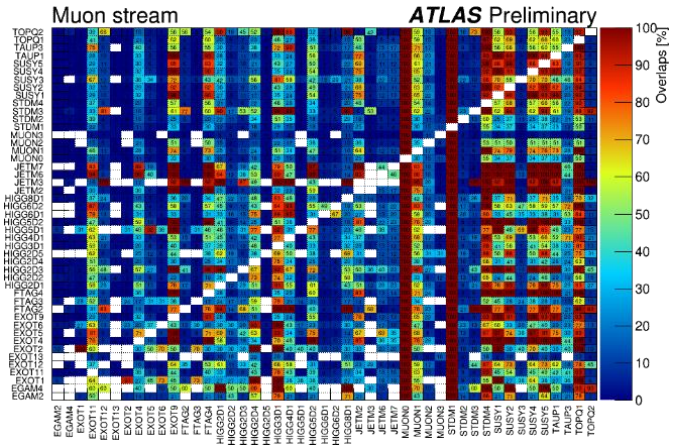
- **EventIndex** это система, созданная с целью каталогизировать события эксперимента **ATLAS,** полученные на установке или в результате компьютерного моделирования

Событие – основная единица данных ATLAS

- Каждое событие содержит информацию о результатах одного столкновения частиц:
  - сигналы с детекторов
  - восстановленные частицы и их параметры
  - триггерные решения
- Каждое событие имеет уникальный идентификатор определяемый номером Run-а и номером события
- Информация о событии хранится в нескольких экземплярах на различных серверах сети GRID
- Форматы в которых представлена информация различаются в зависимости от назначения, например для различных физических анализов создаются наборы данных (**датасеты**) со специально отобранными событиями и их параметрами

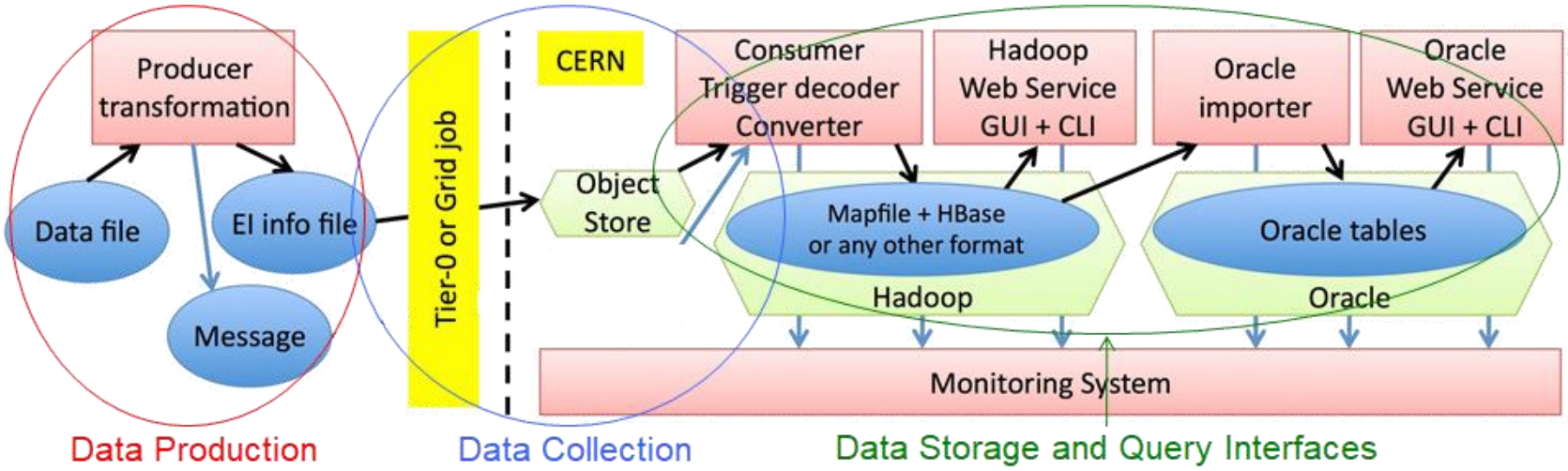
Варианты использования

- **Отбрать событие (Event Picking)**
  - получить событие в нужном формате и обработке
- **Подсчитать и выбрать события на основании триггерных решений**
- **Проверить целостность и согласованность данных**
  - Искажение данных, потерянные и/или повторяющиеся события
- **Создание матриц перекрытия триггерных цепочек**
- **Создание матриц перекрытия наборов данных**
- **Быстрый просмотр датасетов**
  - Поиск требуемых датасетов
  - Составление отчетов
  - Проверка датасетов



Подробнее здесь:
https://twiki.cern.ch/twiki/bin/view/AtlasComputing/EventIndexUseCases

Архитектура EventIndex



Partitioned architecture, определяемая потоком данных

Data production

- Извлечение метаданных событий из файлов, созданных на **Tier-0** или **Grid**

Data collection

- Передача собранной задачами **EI** информации на сервера в **CERN**

Data storage and Query interfaces

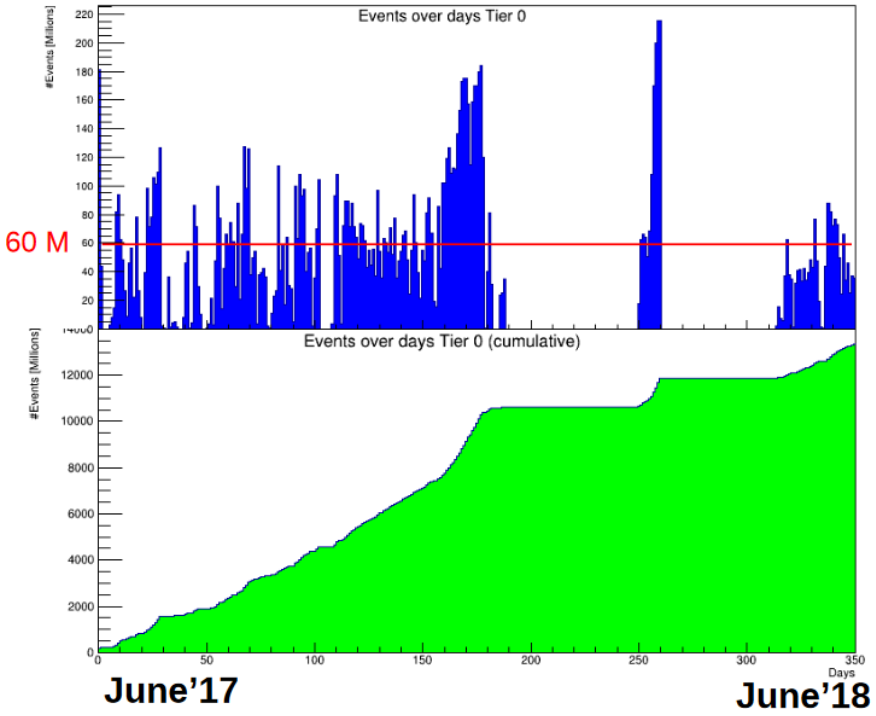
- Обеспечивает постоянное хранение данных **EventIndex.**
- Полная информация хранится в системе **Hadoop**
- Сокращенная информация (без триггеров) хранится в **Oracle,** для быстрых запросов

Monitoring

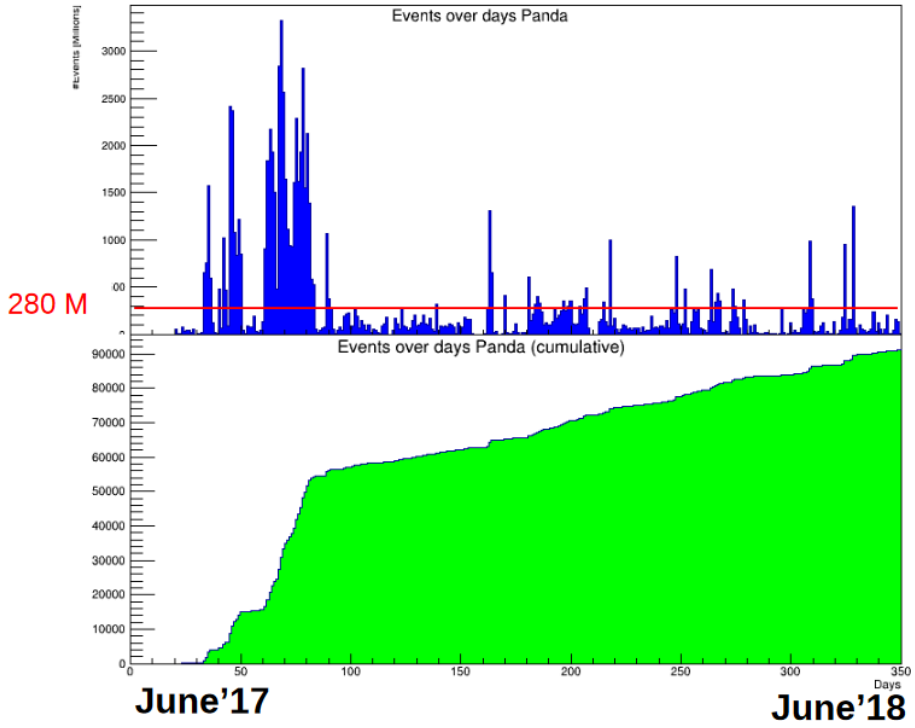
- Позволяет отслеживать работоспособность системы, контролировать потоки данных и выявлять проблемы

Производительность системы

TIER0 @ CERN



GRID



Tier0 @ CERN : ~60 M событий/день

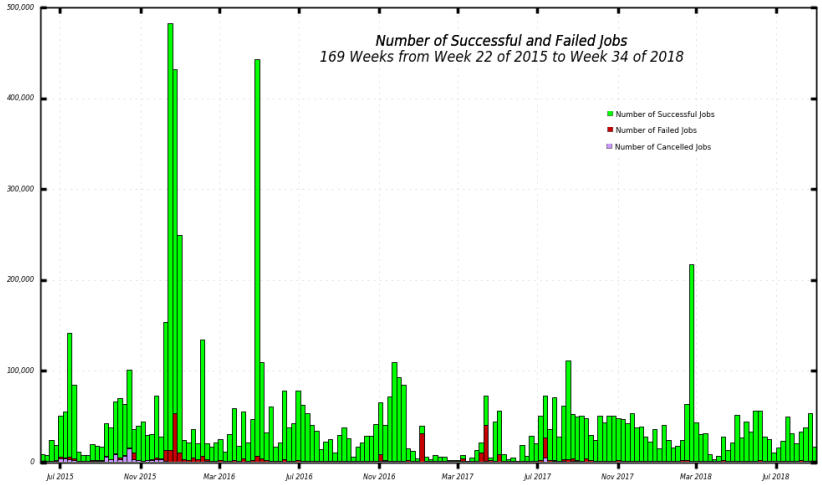
Grid Sites : ~280 M событий/день

Всего **(Tier0 + Grid):**

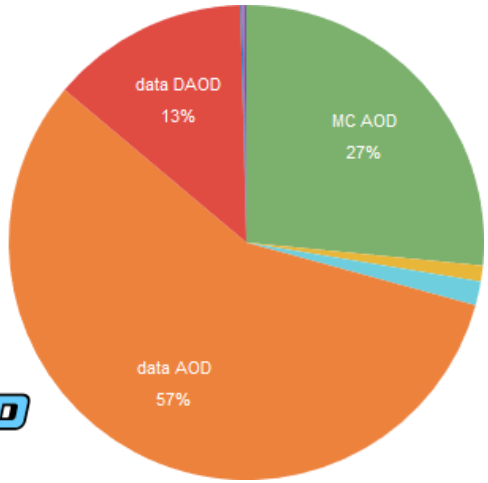
- В пиках индексируется до **3500 M** событий в день
- **~100 Млрд** событий проиндексировано за последний год

Data Production

- На **Tier-0** индексируются полученные на установке данные в формате **AOD,** а также собираются ссылки на сырые данные
- **Данные прошедшие дальнейшую обработку на серверах Grid, обрабатываются на них же. Информацию о готовности данных предоставляют системы AMI и Rucio**
- Периодически поступают на обработку специальные наборы данных для отдельных задач и рабочих групп
  - В последнее время **EI** активно используется для контроля качеств новых
- Система запущена весной **2015** года и работает в штатном режиме
- Сбои крайне редки и вызваны в основном проблемами серверов **Grid** либо поврежденными данными
- Объемы данных в **Hadoop**:
  - Реальные данные: **21** ТБ
  - Монте-Карло: **5** ТБ
  - Другое (бэкап, временные, **etc**…): **150** ТБ

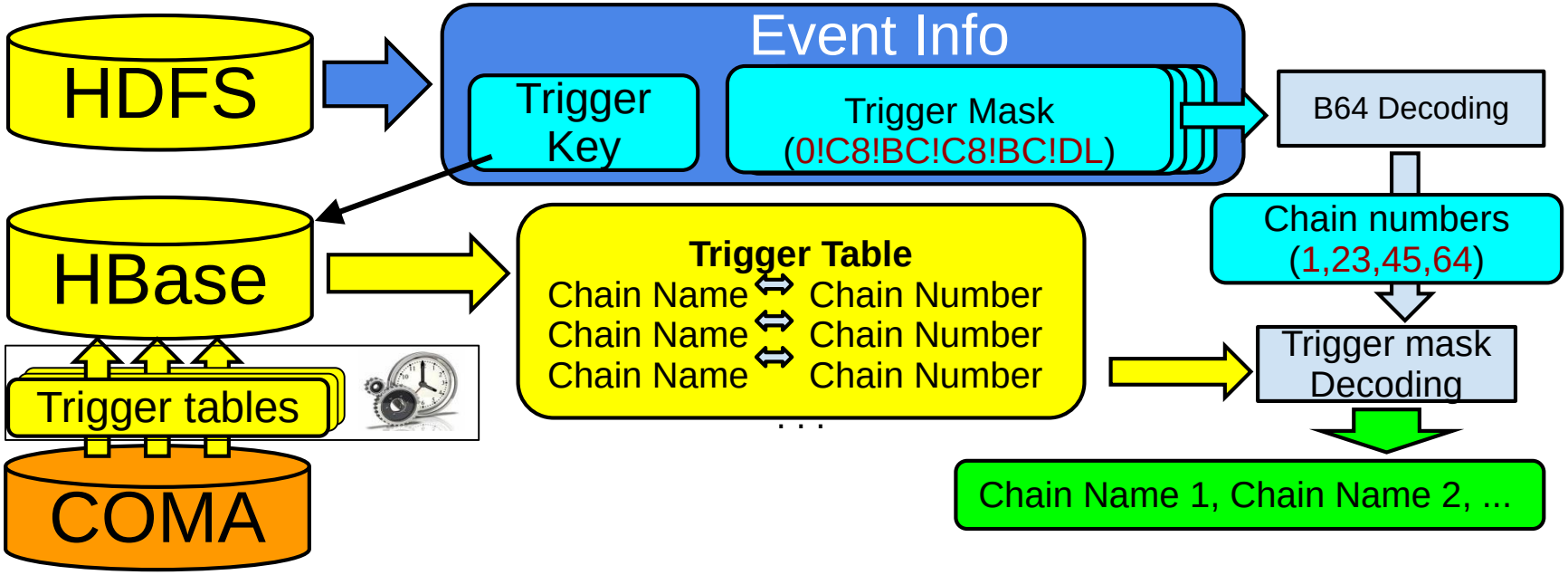


	current
Real AOD	19.2 TB
Real DAOD	4.46 TB
Real DESDM	40 GB
Real DRAW	39 GB
Real ESD	57 GB
Real RAW	0 GB
MC AOD	8.97 TB
MC DAOD	355 GB
MC EVNT	552 GB
Other	9 GB

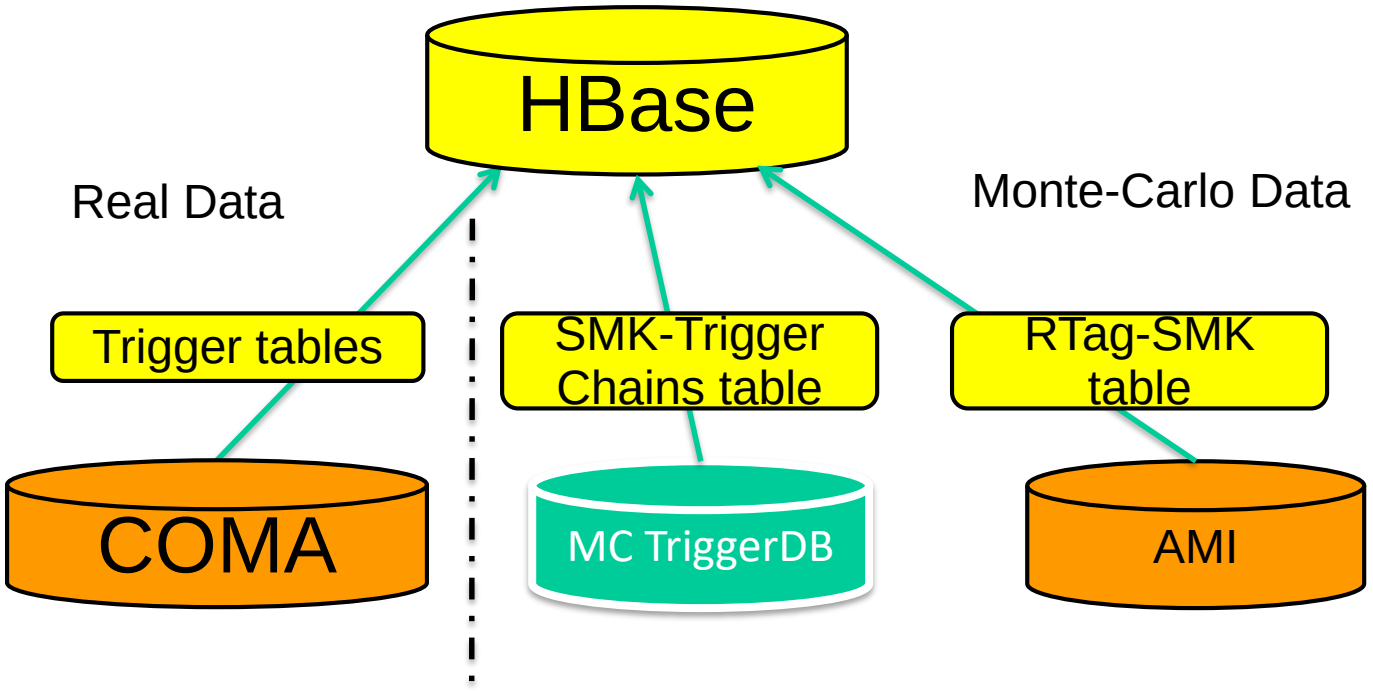


Декодирование триггерной информации

- Триггерные решения записаны в файлах данных в виде битовых полей **(trigger masks).**
- Положение установленных битов **(bit counter)** указывает**,** какие из триггерных цепочек сработали в данном событии**.**
- Триггерная информация сохраняется в **EventIndex** в декодированном виде
- Для данных с установки триггерные биты конверторуются в имена триггерных цепочек используя таблицы соответствий из базы метаданных параметров работы **(COMA, the conditions metadata database)**
- Декодирование триггерной информации для моделированных данных **(MC)** призводится с помощью таблиц взятых из другого источника **(TRIGGERDBMC)**



Триггерная информация для Monte-Carlo



- Если информация о SMK отсутствует в записи события – используется версия реконструкции (**r-tag**)
- В AMI содержится информация о SMK для каждого **r-tag**
- **r-tag** берется из записи события или имени датасета

mc16_13TeV.364109.Sherpa_221_NNPDF30NNLO_Zmumu_MAXHTPTV280_500_CVetoBVeto.merge.AOD.e5271_e5984_s3126_r10517_r10210

Использование триггерной информации

EventIndex позволяет подсчитывать частоту срабатывания триггерных цепочек для оценки эффективности триггеров.

PNG histogram

data18_13TeV.00349842.physics_Main.merge.AOD.f933_m1960 @00	evt=1.8657944E7
L1_EM22VHI	9139599 (48.99%)
L1_EM24VHI	8681555 (46.53%)
L1_EM24VHIM	8477955 (45.44%)
L1_MU20	4951122 (26.54%)
L1_MU20_FTK	4951122 (26.54%)
L1_MU21	4871574 (26.11%)
L1_TAU60	4592084 (24.61%)
L1_J100	4511415 (24.18%)
L1_EM15VHI_TAU40_2TAU15	4254325 (22.80%)
L1_SC111-CJ15	4185692 (22.43%)
L1_HT190-J15s5.ETA21	3824623 (20.50%)
L1_XE50	3812247 (20.43%)
L1_J120	3628486 (19.45%)

PNG histogram

data18_13TeV.00349842.physics_Main.merge.AOD.f933_m1960 @00	evt=1.8657944E7
HLT_j0_perf_ds1_L1J100	4507426 (24.16%)
HLT_e26_lhtight_nod0	2579342 (13.82%)
HLT_mu26_ivarmedium	2520177 (13.51%)
HLT_mu26_ivartight	2456332 (13.17%)
HLT_e26_lhtight_nod0_ivarloose	2327194 (12.47%)
HLT_mu28_ivarmedium	2215895 (11.88%)
HLT_mu28_ivartight	2168251 (11.62%)
HLT_e28_lhtight_nod0	2166217 (11.61%)
HLT_e28_lhtight_nod0_ivarloose_L1EM22VHI	2122560 (11.38%)
HLT_e28_lhtight_nod0_ivarloose	1995290 (10.69%)
HLT_e28_lhtight_nod0_ivarloose_L1EM24VHIM	1935065 (10.37%)
HLT_e32_lhtight_nod0_ivarloose	1627066 (8.72%)
HLT_mu50	807739 (4.33%)
HLT_tau35_medium1_tracktwoEF_tau25_medium1_tracktwoEF_L1DR-TAU20ITAU12I-J25	684287 (3.67%)
HLT_tau35_medium1_tracktwo_tau25_medium1_tracktwo_L1DR-TAU20ITAU12I-J25	628693 (3.37%)

data18_13TeV.00349842.physics_Main.merge.AOD.f933_m1960
* Catalog - Dataset Overlaps - Trigger Statistics - Trigger Overlaps - TagFile Sample - TagFile Info(*) - Journal(run)(tag)(*) - AMI(+) - Rucio(+) - COMA(+)
* Generic: Catalog - Event Index
* For experts: EI - EL - TI - Inspect - Journal - Full Service-oriented Portal

(*) ... may be slow, (+) ... external service

Использование триггерной информации

Матрицы перекрытий используются для оптимизации триггерных меню, доступно также представление в виде графов отношений

PNG

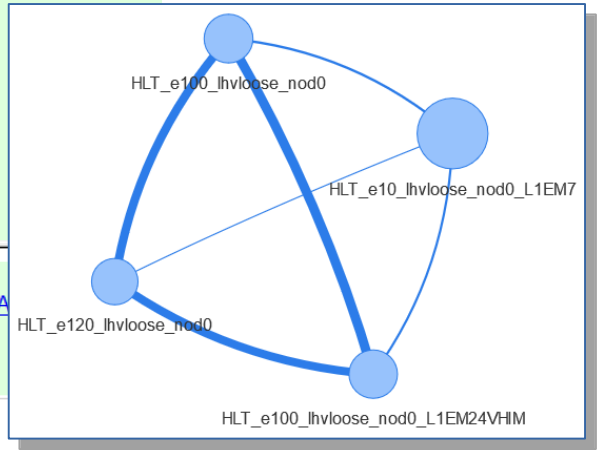
data18_13TeV.00349842.physics_Main.merge.AOD.f933_m1960
evt=18657944,t3=1.8675657E7

Show Selection ☐ inclusive

HLT_e100_lhvloose_nod0	7910568 (42.40%)	☑
HLT_e100_lhvloose_nod0_L1EM24VHIM	7798626 (41.80%)	☑
HLT_e10_lhvloose_nod0_L1EM7	14813658 (79.40%)	☑
HLT_e120_lhvloose_nod0	7164288 (38.40%)	☑
HLT_e120_lhvloose_nod0_L1EM24VHIM	7556085 (40.50%)	☐
HLT_e12_lhloose_cutd0dphideta_L1EM10VH	11865852 (63.60%)	☐
HLT_e12_lhloose_nod0		
HLT_e12_lhloose_nod0_L1EM10VH		
HLT_e12_lhloose_nodeta_L1EM10VH		
HLT_e12_lhloose_nodphires_L1EM10		
HLT_e12_lhmedium_nod0		
HLT_e12_lhvloose_nod0_L1EM10VH		
HLT_e13_etcut_trkcut_xs30_j15_perf		
HLT_e13_etcut_trkcut_xs30_j15_perf		
HLT_e13_etcut_trkcut_xs30_j15_perf		
HLT_e13_etcut_trkcut_xs30_j15_perf		
HLT_e13_etcut_trkcut_xs30_xe30_mt		
HLT_e14_lhtight_nod0	13638267 (73.10%)	☐
HLT_e15_lhvloose_nod0_L1EM7	14757687 (79.10%)	☐
HLT_e17_lhloose_cutd0dphideta	10653147 (57.10%)	☐
HLT_e17_lhloose_cutd0dphideta_L1EM15	13041243 (69.90%)	☐
HLT_e17_lhloose_nod0	10653147 (57.10%)	☐
HLT_e17_lhloose_nod0_L1EM15	13041243 (69.90%)	☐
HLT_e17_lhloose_nodeta_L1EM15	13041243 (69.90%)	☐

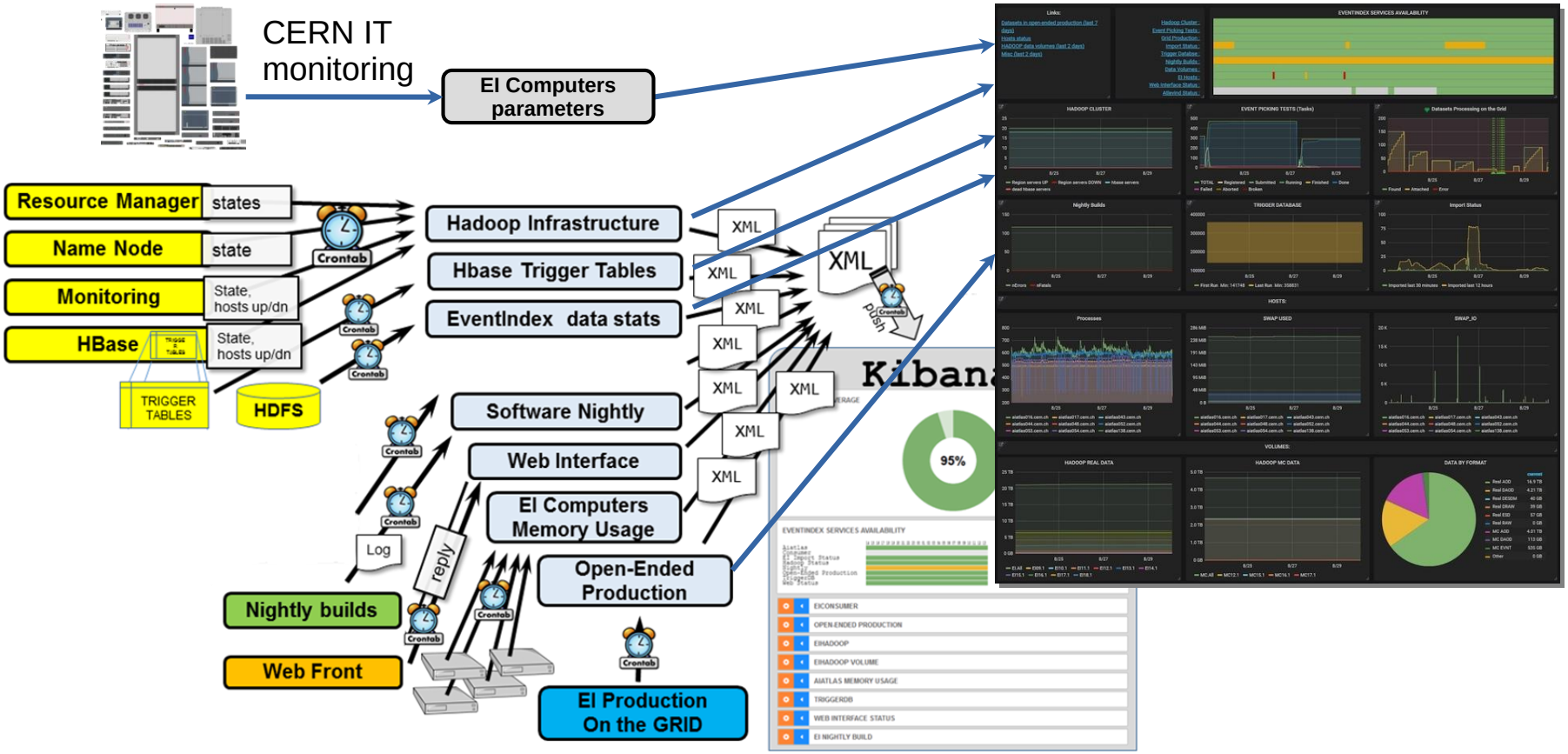
data18_13TeV.00349842.physics_Main.merge.AOD.f933_m1960 evt=18657944,t3=1.8675657E7	HLT_e100_lhvloose_nod0 **7910568** (42.40%)	HLT_e100_lhvloose_nod0_L1EM24VHIM **7798626** (41.80%)	HLT_e10_lhvloose_nod0_L1EM7 **14813658** (79.40%)	HLT_e120_lhvloose_nod0 **7164288** (38.40%)
HLT_e100_lhvloose_nod0 **7910568** (42.40%)	**7910568** (42.40%) (7910568) 100.00%,100.00%	**7462800** (40.00%) (8246394) 94.34%,95.69%	**7910568** (42.40%) (14813658) 100.00%,53.40%	**6903090** (37.00%) (8171766) 87.26%,96.35%
HLT_e100_lhvloose_nod0_L1EM24VHIM **7798626** (41.80%)	**7462800** (40.00%) (8246394) 95.69%,94.34%	**7798626** (41.80%) (7798626) 100.00%,100.00%	**7798626** (41.80%) (14813658) 100.00%,52.64%	**6791148** (36.40%) (8171766) 87.08%,94.79%
HLT_e10_lhvloose_nod0_L1EM7 **14813658** (79.40%)	**7910568** (42.40%) (14813658) 53.40%,100.00%	**7798626** (41.80%) (14813658) 52.64%,100.00%	**14813658** (79.40%) (14813658) 100.00%,100.00%	**7164288** (38.40%) (14813658) 48.36%,100.00%
HLT_e120_lhvloose_nod0 **7164288** (38.40%)	**6903090** (37.00%) (8171766) 96.35%,87.26%	**6791148** (36.40%) (8171766) 94.79%,87.08%	**7164288** (38.40%) (14813658) 100.00%,48.36%	**7164288** (38.40%) (7164288) 100.00%,100.00%

data18_13TeV.00349842.physics_Main.merge.AOD.f933_m1960
* Catalog - Dataset Overlaps - Trigger Statistics - Trigger Overlaps - TagFile Sample - TagFile Info(*) - Journal(run)(tag)(*) - A
* Generic: Catalog - Event Index
* For experts: EI - EL - TI - Inspect - Journal - Full Service-oriented Portal
(*) ... may be slow, (+) ... external service



Мониторинг EventIndex с использованием Grafana

- **Данные о состоянии и производительности системы(~15k** значений в день) **собираются разными способами (сканирование журналов и HDFS, REST запросы и анализ WEB-страниц)**
- **Сбор и обработка информации управляется планировщиком**
- **Результаты загружаются в InfluxDB / Grafana через REST**



Панель EventIndex в Grafana

В настоящее время:

- Информация о событиях, относящаяся к различным этапам обработки (RAW, ESD, AOD, DAOD) физически хранится в разных файлах системы хранения данных HADOOP (HDFS)

В перспективе:

- Одна единственная логическая запись для события
  - Идентификация события
  - Неизменная информация (lumiblock, триггер)
  - Для каждого этапа обработки:
    - Ссылка на алгоритм (конфигурация задачи обработки)
    - Указатели на результаты обработки
    - Флаги для оффлайн селекции
- Виртуальный датасеты – логические набор событий
  - Создается явным образом из коллекции идентификаторов событий, либо на основании выборки из уже имеющихся наборов

В настоящее время:

- Все процессы ATLAS производят ~30 миллиардов событий в год (в среднем до 350 Гц).
- Это новые и обновленные данные, относящиеся ко всем годам, полученные как на установке, так и в результате моделирования
- EventIndex обрабатывает 8 миллионов файлов в день

В перспективе:

- Необходимо масштабировать систему для ожидаемого роста потоков данных :
  - Run3 (2021-2023): 35 млрд новых реальных событий в год и 100 млрд. моделированных событий в год (рост как минимум на половину порядка величины).
  - Run4: 100 млрд новых реальных событий в и 300 млрд. Монте-Карло событий в год.
  - Плюс репликация и репроцессинг

Развитие EventIndex

- Используемые в EventIndex технологии в основном позволили обеспечить производительность удовлетворительную для условий Run 2.
- С ростам объема данных стали проявлятся проблемы масштабиования:
  - Медленная обработка запросов, рост объема данных
- В Run3 ожидается существенное возрастание объемов данных т темпа их поступления
- Предполагается провести модернизацию EventIndex с использованием современных технологий BigData для работы в условиях Run 3
- Модернизация предполагает использование **HBase** для хранения данных и **Apache Phoenix** для организации запросов.
- Эти продукты принадлежат к семейству Hadoop и будут поддерживаться CERN на протяжении всего Run 3.

Развитие EventIndex. HBase

- Apache HBase принадлежит к семейству HADOOP.
  - Нереляционная база данных с открытым кодом
  - Распределенное и масштабирование хранилище данных
- HBase организует данные в виде таблиц
  - Ряды таблицы имеют уникальный ключ-иднтификатор
  - Каждй ряд может иметь свою схему
  - Данные в каждом ряду могут быть сгруппированы в заранее заданное семейства колонок
  - Доступ к значениям возможен по ключу и имени колонки
- Свойства такой организации (**RowKey**):
  - Относительно небольшой размер
  - Позволяет уникально идентифицировать события
  - Позволяет производить поиск по интервалам
  - При росте таблиц используется гомогенное пространство RowKey
- Hbase уже использовалась для триггерных таблиц

- Средства для SQL на HBase:
  - **Apache Impala**
  - **Apache Hive**
  - **Apache Spark**

  Обращение с отображением ключей рядов должно осуществляется приложением преимущественно для пакетных задач

Apache Phoenix

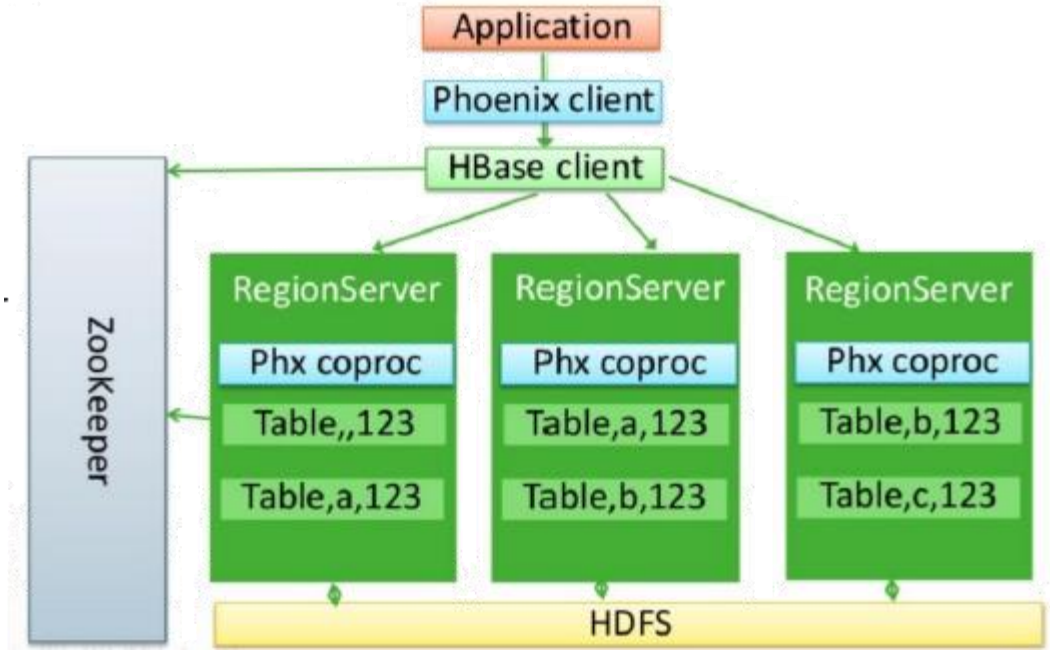
Транзакционная система и операционная аналитика для HBase

- Принимает SQL запросы
- Преобразует их в серии сканирований в HBase
- Напрямую использует API HBase, а также сопроцессоры и сспециальные фильтры
- Выдает результаты запросов JDBC

- Дизайн RowKey адаптируется к типам и размерам Phoenix за счет небольшого уменьшения производительности
- Phoenix позволяет использовать поля RowKey в запросах, при этом они сохраняются как один объект в HBase

- Запросы Phoenix к HBase были протестированы на прототипах таблиц EventIndex с обнадеживающими результатами

- Варианты схем таблиц
- Готовы базовые функции
- Идет работа над производительностью и интерфейсами
- Необходимо продолжать тестирование с большими объёмами данных для получения показателей производительности



- Планируется ввести новую систему в действие в течении **2019** года, параллельно с уже имеющейся
- Снятие с эксплуатации старой системы планируется в течении **2020** г (до начала Run 3 на LHC)

- Технологии работы с «Большими Данными» активно развиваются
- Появления и развитие новых средств обработки и хранения данных позволяет провести **глубокую модернизацию** EventIndex с переходом на принципиально другую платформу и **расширением возможностей**
- В любом случае, участие в проекте **EventIndex** позволяет не только внести вклад в эксперимент ATLAS, но и позволяет работать с **самым современным** технологиями **Больших Данных**, с перспективой использования полученного опыта в других экспериментах или образовательных проектах.