**Hadoop**

**Руководство администратора**

Оглавление

[1 История изменений 4](#_Toc522275100)

[2 Глоссарий 5](#_Toc522275101)

[3 Ссылки 7](#_Toc522275102)

[4 Общие сведения о системе 8](#_Toc522275103)

[4.1 Источники и приёмники данных 8](#_Toc522275104)

[4.2 Структура домашнего каталога в HDFS 8](#_Toc522275105)

[5 Требования к аппаратному и программному обеспечению 9](#_Toc522275106)

[5.1 Требования к аппаратному обеспечению 9](#_Toc522275107)

[5.2 Требования к программному обеспечению 9](#_Toc522275108)

[6 Инструкции по инсталляции и деинсталляции системы 10](#_Toc522275109)

[6.1 Инсталляция 10](#_Toc522275110)

[6.1.1 Настройка кластера BDA 10](#_Toc522275111)

[6.1.2 Настройка HDFS 10](#_Toc522275112)

[6.1.3 Подготовка БД Hive 12](#_Toc522275113)

[6.1.4 Установка и настройка Kerberos 12](#_Toc522275114)

[6.1.5 Настройка HUE Web UI 12](#_Toc522275115)

[6.1.6 Установка и настройка Oozie Workflows 13](#_Toc522275116)

[6.1.6.1 Подготовка домашнего каталога UNIX 13](#_Toc522275117)

[6.1.6.2 Подготовка домашнего каталога HDFS 13](#_Toc522275118)

[6.1.6.3 Установка патчей с Oozie Workflows 13](#_Toc522275119)

[6.1.7 Перечень Oozie Worflows и их параметров 14](#_Toc522275120)

[6.1.8 Настройка расписания и запуск регламентных Oozie Worflows 14](#_Toc522275121)

[6.1.9 Ручной запуск Oozie Worflows 14](#_Toc522275122)

[6.2 Деинсталляция 15](#_Toc522275123)

[7 ETL-процессы 16](#_Toc522275124)

[7.1 Регулярные загрузки 16](#_Toc522275125)

[7.2 Отбор и проверка данных для загрузок 16](#_Toc522275126)

[7.2.1 Ручной режим запуска 16](#_Toc522275127)

[7.2.2 Автоматический режим запуска 16](#_Toc522275128)

[7.3 Запуск загрузок 16](#_Toc522275129)

[7.4 Загрузка исторических данных 16](#_Toc522275130)

[8 Мониторинг работы системы 17](#_Toc522275131)

[8.1 Мониторинг журнала работы workflow 17](#_Toc522275132)

[9 Логирование 22](#_Toc522275133)

# История изменений

|  |  |  |  |
| --- | --- | --- | --- |
| **Дата** | **Номер версии** | **ФИО вносившего корректировки** | **Описание корректировок** |
| 09.07.2018 | 0.1 | Коротков Р.Е. | Документ создан |
| 13.07.2018 | 0.2 | Шилин В.А. | Добавлено описание oozie workflows. 6.1.6, 6.1.7, 6.1.8, 6.1.9 |
| 17.08.2018 | 0.3 | Коротков Р.Е. | Актуализация 7го пункта |
| 31.08.2018 | 0.4 | Назаренко Р | Актуализация ручного перезапуска упавших джобов |
| 12.09.2018 | 0.5 | Шилин В.А. | Актуализация 7го пункта. Добавлены etl для загрузки ОперБлока из ods |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

# Глоссарий

| *Понятие* | *Сокращение* | *Определение понятия* |
| --- | --- | --- |
| Apache Hive | Hive | Комплекс программных средств для доступа и управлениями большими объёмами данных в распределённых хранилищах данных |
| DetailData / Область детальных данных |  | Область хранения детальных данных системы Hadoop |
| Hadoop |  | Технология, основанная на распределенной обработке больших объемов данных. Характеризируется доступностью, надежностью, эффективной масштабируемостью, простотой |
| Hadoop Distributed File System | HDFS | Файловая система, предназначенная для хранения файлов больших размеров, поблочно распределённых между узлами вычислительного кластера |
| Hadoop | Hadoop | Модуль системы для обработки и хранения слабоструктурированных исторических данных |
| Relational | R | Система для обработки и хранения реляционных данных |
| SourceData/ SourceFiles/ Область исходных данных/ Область исходных файлов |  | Область хранения исходных файлов данных системы Hadoop |
| Альфа-банк | Банк | Альфа-банк |
| Бизнес-дата/ Отчётная дата (<отчётная дата>) |  | дата, за которую осуществляется. Физически это значение поля DAY из Hive-таблиц DetailData-слоя |
| Система источник | СИ | Система-источник для области хранения исходных файлов |
| Система |  | * ПО, включающее в себя: * Загрузки (Oozie workflow) и требуемые для их функционирования библиотеки; * Домашний каталог системы и все включенные в него файлы; * Все процессы, выполняемые на кластере, инициированные загрузкой или пользователем Системы; |
| Oracle Big Data Appliance | BDA, кластер | Программно-аппаратный комплекс Oracle, содержащий несколько узлов (физических машин) объединённых в кластер, и содержащий комплект ПО Cloudera для работы с Hadoop. |
| Cloudera Distribution Including Apache Hadoop | CDH | Комплект ПО Cloudera для работы с Hadoop |
| Apache Oozie | Oozie | Инструментарий для организации и планирования потоков заданий для Apache Hadoop. |
| Apache Hive | Hive | SQL-движок, построенный поверх архитектуры Apache Hadoop, обладающий большинством возможностей современного SQL-языка. |
| Cloudera HUE | HUE | Web GUI инструментарий, позволяющий удобно работать с компонетами Apache Hadoop, такими как Hive, Oozie, HDFS и другими, а так же обладающий возможностями аутентификации пользователей. |
| Apache Hadoop Distributed File System | HDFS | Распределенная отказоустойчивая файловая система Apache Hadoop. |
| Apache Oozie Workflow | workflow | Задача, выполняемая средствами Apache Oozie. Может содержать другие подзадачи, напрмер java или hive, а также другие workflow. |
| Oozie action | action | Подзадача внутри Oozie workflow. |
| Hive User Defined Functions | UDF | Специально разработанные функции, которые можно использовать в SQL-запросах Hive. |

# Ссылки

| *№* | *Наименование документа* | *Номер версии* | *Дата создания версии* | *Место нахождения документа/ссылка* |
| --- | --- | --- | --- | --- |
|  | DWH patch rules | 2.1.13 |  | [**http://confluence.moscow.alfaintra.net/pages/viewpage.action?pageId=196800792**](http://confluence.moscow.alfaintra.net/pages/viewpage.action?pageId=196800792) |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

# Общие сведения о системе

## Источники и приёмники данных

## Структура домашнего каталога в HDFS

# Требования к аппаратному и программному обеспечению

## Требования к аппаратному обеспечению

## Требования к программному обеспечению

# Инструкции по инсталляции и деинсталляции системы

## Инсталляция

### Настройка кластера BDA

Требуется установка следующих сервисов Cloudera (в случае их отсутствия в Cloudera Manager):

* HUE;
* Oozie;
* Hive;

Для каждого пользователя Системы необходимо завести учетную запись Oracle Linux. Требуется только наличие учетной записи, поэтому они могут быть не активны (отключены). В случае использования LDAP, заведение учетных записей не требуется.

В случае использования ПО Kerberos, необходимо завести для каждого пользователя соответствующего principal, имя которого соответствует имени учетной записи Oracle Linux.

### Настройка HDFS

В Linux на последней ноде кластера:  
  
Добавить группы для технического пользователя, без них не будет работать создание external таблиц на данных в каталоге пользователя:  
  
**dcli -C groupadd tech\_oozie\_w**  
**dcli -C groupadd tech\_oozie\_r**  
  
Добавить пользователя и нужные группы. Основная группа обязательна, дополнительные группы могут варьироваться.  
**dcli -C adduser -g tech\_oozie\_w -G s\_dwh\_w,s\_dwh\_tmp\_w,s\_dmacrm\_w,s\_dmacrm\_tmp\_w,s\_dwh4dm\_w,s\_dwh4dm\_tmp\_w,s\_sourcedata\_w,l\_deriveddata\_w tech\_oozie**  
  
При желании можно проверить что пользователь создался верно.  
**dcli -C id tech\_oozie**  
  
Задать пароль:  
**passwd tech\_oozie**  
  
В Linux на ноде «текущий active kerberos» добавить принципала:  
(Если пользователь уже есть на каком либо кластере - действие не требуется)  
  
**kadmin.local**  
**> add\_principal tech\_oozie**   
  
В beeline из под суперпользователя (hdfs) добавляем привилегии:  
  
**CREATE ROLE tech\_oozie\_w;**  
**GRANT ALL ON URI 'hdfs://bda21/user/tech\_oozie' TO ROLE tech\_oozie\_w;**  
  
**CREATE ROLE tech\_oozie\_r;**  
**GRANT SELECT ON DATABASE tech\_oozie TO ROLE tech\_oozie\_r;**  
  
**GRANT ROLE tech\_oozie\_w TO GROUP tech\_oozie\_w;**  
**GRANT ROLE tech\_oozie\_r TO GROUP tech\_oozie\_r;**  
  
В Linux на последней ноде кластера:  
Раздать права на каталог, описываем ACL, задаём лимит на пространство:  
  
**hdfs dfs -chown tech\_oozie:tech\_oozie\_w /user/tech\_oozie**  
**hdfs dfs -chmod 770 /user/tech\_oozie**  
**hdfs dfs -setfacl --set user::rwx,group::rwx,group:tech\_oozie\_r:r-x,other::--- /user/tech\_oozie**  
**hdfs dfs -setfacl --set default:user::rwx,default:group::rwx,default:group:tech\_oozie\_r:r-x,default:other::--- /user/tech\_oozie**  
**hdfs dfsadmin -setSpaceQuota 53687091200 /user/tech\_oozie**  
**hadoop fs -count -q /user/tech\_oozie**

### Подготовка БД Hive

### Установка и настройка Kerberos

1. Установить и настроить Kerberos KDC на один из узлов кластера (любой) согласно официальной инструкции MIT/Cloudera;
2. Установить клиентские библиотеки Kerberos на каждый узел кластера, настроить krb5.conf согласно официальной инструкции MIT/Cloudera;
3. Настроить Kerberos на кластере DBA средствами Cloudera Manager согласно официальной инструкции Oracle/Cloudera;
4. Подготовить Kerberos keytab-файл (необходим для работы Hive JDBC):
   * Создать keytab-файл:
     + На узле с Kerberos KDC выполнить:
     + >kadmin.local;
     + Создать keytab (Роль - developer):
     + >kadmin: xst -norandkey -k < **tech\_oozie** >.keytab < **tech\_oozie** >/\_HOST@realm
     + (где < **tech\_oozie** >– имя пользователя с пользователя Kerberos/Linux, \_HOST – полное имя хоста, на котором работает сервис Oozie, realm – Kerberos realm).
   * Поместить сформированный файл < **tech\_oozie** >.keytab в каталог Системы *<hdfs\_home>*/config/
   * *Примечание*: Необходимо выполнять данный шаг каждый раз при изменении пароля Kerberos.
   * Пользователю необходимы права на доступ к hdfs в зависимости от проекта (например вхождение в группы s\_dwh\_w, s\_sourcedata\_w для доступа к s\_dwh и s\_sourcedata соответственно)

### Настройка HUE Web UI

1. Перейти по адресу http://<hue-server>:8888/ (где <hue-server> – адрес узла кластера, на котором работает HUE);
2. Выполнить создание необходимых ролей и пользователей (**tech\_oozie**) с указанием опции создания домашней директории.
3. Развернуть необходимые workflow, описанные в следующем разделе.

### Установка и настройка Oozie Workflows

#### Подготовка домашнего каталога UNIX

* Создать каталог для патчей в Unix, если его еще нет. Под пользователем <TECH\_OOZIE> выполнить следующие команды:
* mkdir /home/${USER}/oozie\_patches
* Создать основной unix-каталог для oozie\_workflows, если его еще нет. Под пользователем <TECH\_OOZIE> выполнить следующие команды:
* mkdir /home/${USER}/oozie\_workflows
* mkdir /home/${USER}/oozie\_workflows/atom
* mkdir /home/${USER}/oozie\_workflows/ctl
* mkdir /home/${USER}/oozie\_workflows/reg
* mkdir /home/${USER}/oozie\_workflows/man

#### Подготовка домашнего каталога HDFS

* Создать директории для внешних таблиц, если их еще нет. Под пользователем <TECH\_OOZIE> выполнить следующие команды:
* hadoop fs -mkdir /user/${USER}/tmp
* Создать директории для oozie workflow, если их еще нет. Под пользователем <TECH\_OOZIE> выполнить следующие команды:
* hadoop fs -mkdir /user/${USER}/oozie\_workflows
* hadoop fs -mkdir /user/${USER}/oozie\_workflows/atom
* hadoop fs -mkdir /user/${USER}/oozie\_workflows/ctl
* hadoop fs -mkdir /user/${USER}/oozie\_workflows/reg
* hadoop fs -mkdir /user/${USER}/oozie\_workflows/man

#### Установка патчей с Oozie Workflows

* Подготовка БД Hive. Запустить hive скрипты, расположенные в патче в каталоге hive\_scripts, в соответствующих БД. Перед запуском необходимо подставить в скрипты корректные параметры в соответсвии с ReleaseNotes.
* Подготовка Oozie Workflows и настройка параметров:
  + Скопировать каталог с патчем в unix под пользователем <TECH\_OOZIE> в специально отведенное для патчей место. Например в /home/<TECH\_OOZIE>/oozie\_patches/<PATCH\_NAME>
  + В unix, в каталоге патча, настроить параметры регламентных потоков (/home/<TECH\_OOZIE>/oozie\_patches/<PATCH\_NAME>/oozie\_workflows/reg). Настройка параметров выполняется в файлах start.sh и job.properties.
  + В unix, в каталоге патча, настроить параметры ручных потоков (/home/<TECH\_OOZIE>/oozie\_patches/<PATCH\_NAME>/oozie\_workflows/man). Настройка параметров выполняется в файлах start.sh и job.properties.
  + Подготовленные каталоги потоков нужно поместить в hdfs (/user/<TECH\_OOZIE>/oozie\_workflows) и в основной unix-каталог для workflows (/home/<TECH\_OOZIE>/oozie\_workflows) . Для этого в каталоге патча есть скрипт (/home/<TECH\_OOZIE>/oozie\_patches/<PATCH\_NAME>/oozie\_workflows/reload.sh.

### Перечень Oozie Worflows и их параметров



### Настройка расписания и запуск регламентных Oozie Worflows

* Каждый регламентный поток (reg) содержит в своем каталоге (/home/<TECH\_OOZIE>/oozie\_workflows/reg/<WF\_NAME>) координатор (coordinator.xml), файл параметров (job.properties) и скрипт запуска (start.sh). Параметры запуска настраиваются в start.sh и job.properties в процессе установки патча.
* Параметры расписания задаются в coordinator.xml:

**<coordinator-app name="crd\_reg\_dmacrm\_daily"**

**frequency="${coord:days(1)}"**

**start="2018-07-03T10:00+0300"**

**end="2999-07-03T10:00+0300"**

**timezone="GMT+03:00"**

**xmlns="uri:oozie:coordinator:0.4">**

**frequency** – периодичность запуска

**start** – дата начала регламентной загрузки. От этой даты будут откладываться интервалы периодичности, описанные в **frequency**.

**end** – дата конца регламентной загрузки.

* Для запуска регулярной загрузки нужно в unix запустить /home/<TECH\_OOZIE>/oozie\_workflows/reg/<WF\_NAME>/start.sh.

### Ручной запуск Oozie Worflows

* Для ручного запуска любого потока нужно создать файл параметров job.properties и скрипт запуска start.sh. Если поток нужно запускать несколько раз подряд за различные временные интервалы, то дополнительно создается и coordinator.xml
* Файлы можно создавать в произвольном unix-каталоге. В рамках патчей все ручные (не регламентные) запуски содержатся в каталоге /home/<TECH\_OOZIE>/oozie\_workflows/man
* Скрипт для ручного запуска оформляется по аналогии с .

В параметре **workflow\_application\_path** необходимо указать полный путь к запускаемому workflow.

Если используется координатор, то в параметре **coord\_application\_path** необходимо указать полный путь к координатору.

* Файл параметров ручного запуска оформляется по аналогии с .

Здесь нужно задать все параметры, используемые в workflow и координаторе (если он есть). Список параметров можно найти в [Перечене Oozie Worflows и их параметров](#_Toc420428404).

Если workflow запускается с координатором, то в job.properties нужно указать:

**oozie.coord.application.path=${coord\_application\_path}**

Если без координатора, то:

**oozie.wf.application.path=${workflow\_application\_path}**

* Файл координатора для ручного запуска оформляется по аналогии с.
  + в тэге coordinator-app настраивается периодичность запуска
  + в тэге <action><workflow><app-path> указывается заданный в start.sh путь к workflow **workflow\_application\_path**
  + в тэге <configuration> передаются параметры workflow. Список параметров можно найти в [Перечене Oozie Worflows и их параметров](#_Toc420428404).

## Деинсталляция

# ETL-процессы

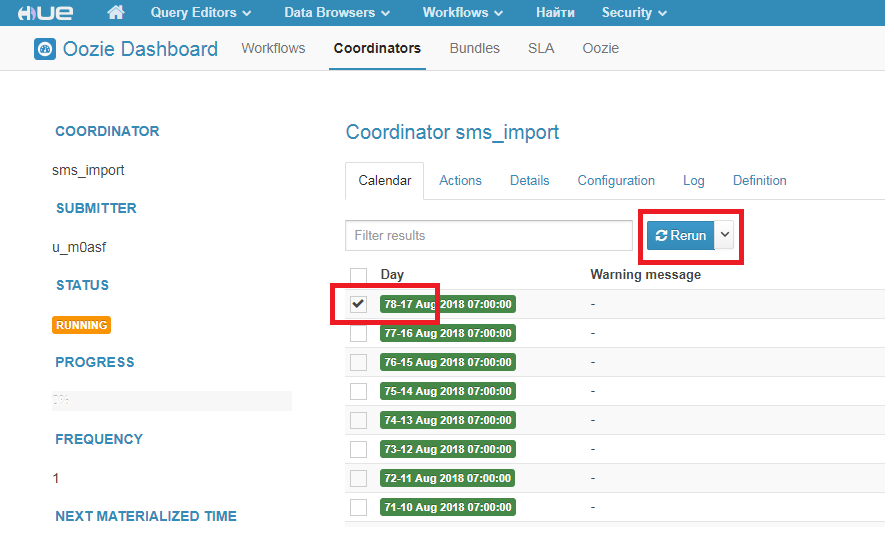
## Регулярные загрузки

|  |  |  |  |
| --- | --- | --- | --- |
| REG поток | CTL поток | Пререквизиты | Примечание |
| wf\_reg\_dmacrm\_daily | wf\_ctl\_dmacrm\_daily\_dictionaries | Нет | Загрузка справочников витрины s\_dmacrm |
| wf\_reg\_dmacrm\_daily | wf\_ctl\_dmacrm\_daily\_facts | Нет | Загрузка фактов витрины s\_dmacrm |
| wf\_reg\_dmacrm\_monthly | wf\_ctl\_dmacrm\_monthly\_day4\_facts | Нет | Загрузка фактов витрины s\_dmacrm |
| wf\_reg\_dmrb\_daily | wf\_ctl\_dmrb\_daily\_dictionaries | Нет | Загрузка справочников витрины s\_dmrb |
| wf\_reg\_dmrb\_daily | wf\_ctl\_dmrb\_daily\_facts | Нет | Загрузка фактов витрины s\_dmrb |
| wf\_reg\_dwh\_daily | wf\_ctl\_dwh\_daily\_dictionaries | Нет | Загрузка справочников витрины s\_dwh |
| wf\_reg\_dwh\_daily | wf\_ctl\_dwh\_daily\_facts | Нет | Загрузка фактов витрины s\_dwh |
| wf\_reg\_dmfr\_daily | wf\_ctl\_dmfr\_daily\_dictionaries | Нет | Загрузка справочников витрины s\_dmfr |
| wf\_reg\_dmfr\_daily | wf\_ctl\_dmfr\_daily\_facts | Нет | Загрузка фактов витрины s\_dmfr |
| wf\_reg\_dmfrua\_daily | wf\_ctl\_dmfrua\_daily\_facts | Нет | Загрузка фактов витрины s\_dmfrua |
| wf\_reg\_dwh4dm\_daily | wf\_ctl\_dwh4dm\_daily\_facts | Нет | Загрузка фактов витрины s\_dwh4dm |
| wf\_reg\_dmpr\_daily | wf\_ctl\_dmpr\_daily\_dictionaries | Нет | Загрузка справочников витрины s\_dmpr |
| wf\_reg\_dmpr\_weekly | wf\_ctl\_dmpr\_weekly\_dictionaries | Нет | Загрузка справочников витрины s\_dmpr |
| wf\_reg\_dmpp\_weekly | wf\_ctl\_dmpp\_weekly\_facts | Нет | Загрузка фактов витрины s\_dmpp |
| wf\_reg\_adb\_daily | wf\_ctl\_adb\_daily\_dictionaries | Нет | Загрузка справочников витрины s\_sourcedata |
| wf\_reg\_adb\_daily | wf\_ctl\_adb\_daily\_facts | Нет | Загрузка фактов витрины s\_sourcedata |
| wf\_reg\_buster\_tpii\_replica\_daily | wf\_ctl\_buster\_tpii\_replica\_daily | Нет | Загрузка фактов витрины s\_sourcedata |
| wf\_reg\_dsacrm\_dwhinform\_daily | wf\_ctl\_dsacrm\_dwhinform\_daily | Нет | Загрузка справочников витрины s\_sourcedata |
| wf\_reg\_odb\_odb\_cm\_daily | wf\_ctl\_odb\_odb\_cm\_daily | Нет | Загрузка справочников витрины s\_sourcedata |
| wf\_reg\_rbprod\_azheleznyakov\_daily | wf\_ctl\_rbprod\_azheleznyakov\_daily | Нет | Загрузка фактов витрины s\_sourcedata |
| wf\_reg\_sandbox\_odh108\_ocrm\_daily | wf\_ctl\_sandbox\_odh108\_ocrm\_daily | Нет | Загрузка справочников витрины s\_sourcedata |
| wf\_reg\_vertica\_public\_daily | wf\_ctl\_vertica\_public\_daily | Нет | Загрузка фактов витрины s\_sourcedata |
| wf\_reg\_ods\_whbranch\_oper\_weekly\_sunday | запускается напрямую атом wf\_ods\_whbranch\_oper | Нет | Загрузка фактов витрины s\_sourcedata |
| wf\_reg\_ods\_whbranch\_oper\_weekly\_monday | запускается напрямую атом wf\_ods\_whbranch\_oper | Нет | Загрузка фактов витрины s\_sourcedata |
| wf\_reg\_dmmlm\_daily | wf\_ctl\_dmmlm\_deriveddata\_daily | Нет | Загрузка витрины l\_deriveddata |
| wf\_reg\_nba\_deriveddata\_daily | wf\_ctl\_nba\_deriveddata\_daily | Нет | Загрузка витрины l\_deriveddata |
| wf\_reg\_nba\_deriveddata\_monthly | wf\_ctl\_nba\_deriveddata\_monthly\_day5 | Нет | Загрузка витрины l\_deriveddata |
| wf\_reg\_nba\_deriveddata\_weekly | wf\_ctl\_nba\_deriveddata\_weekly | Нет | Загрузка витрины l\_deriveddata |

## Отбор и проверка данных для загрузок

### Ручной режим запуска

Для ручного перезапуска координатора в hue выбрать перезагружаемые периоды и нажать кнопку Rerun



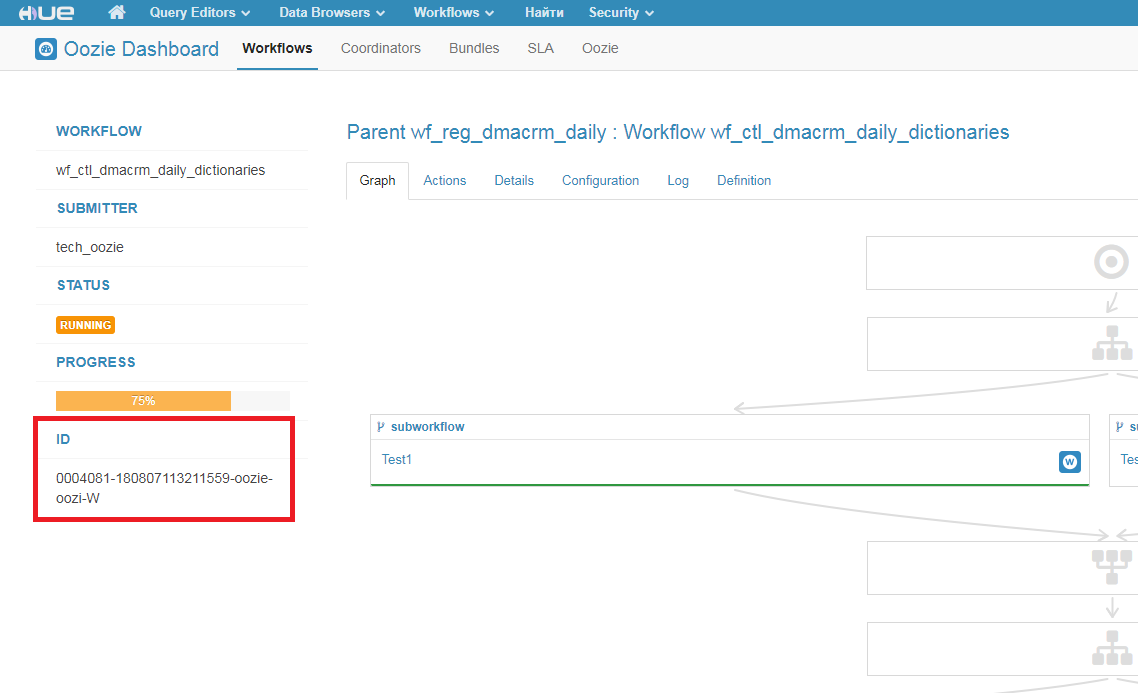
Для запуска в открывшемся окне проверить нажать кнопку Submit

Для ручного перезапуска workflow

(На текущей версии hue через web интерфейс не работает исправлено в новых версиях)

Получить ID задачи по запуску workflow

Для этого зайти в workflow через hue Workflows:



В консоли на кластере выполнить команды:

oozie job -oozie http://bda11node04.moscow.alfaintra.net:11000/oozie -configcontent **[WORKFLOW\_JOB\_ID]** > ./oouzie\_config.xml

В сгенерившемся xml удалить следующий параметры:

<property>

    <name>oozie.coord.application.path</name>

    <value>/user/tech\_oozie/oozie\_workflows/reg/wf\_reg\_dmrb\_daily</value>

  </property>

  <property>

    <name>workflow\_application\_path</name>

    <value>/user/tech\_oozie/oozie\_workflows/reg/wf\_reg\_dmrb\_daily</value>

  </property>

Сохранить и выполнить вот эту команду:

oozie job -oozie http://bda11node04.moscow.alfaintra.net:11000/oozie -rerun **[WORKFLOW\_JOB\_ID]** -Doozie.wf.rerun.failnodes=false -config oouzie\_config.xml

### Автоматический режим запуска

В автоматическом режиме рег потоки устанавливаются на расписание при установке согласно регламенту.

## Запуск загрузок

## Загрузка исторических данных

Для загрузки исторических данных перейти в папку соответствующего man потока на кластере, отредактировать скрипт start.sh заполнив даты загрузки(coord\_start\_time и coord\_end\_time)

Проверить корректность настроек доступов в файле job.properties

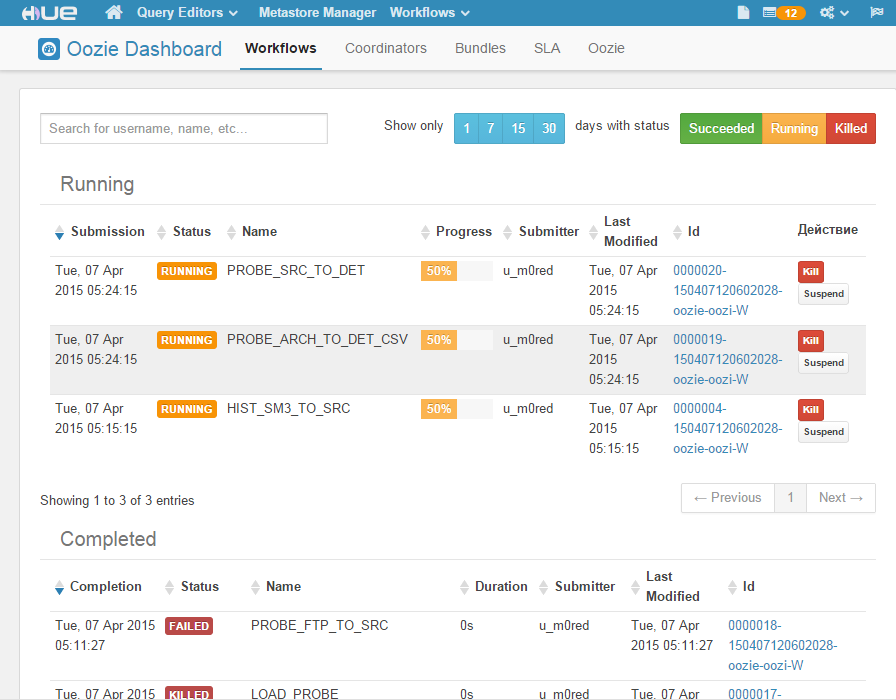
Запустить скрипт start.sh

|  |  |
| --- | --- |
| MAN поток | Примечание |
| crd\_hist\_dmacrm\_bureaureqclient\_sstat | Загрузка s\_dmacrm.bureaureqclient\_sstat |
| crd\_hist\_dmacrm\_bureaureqloan\_sstat | Загрузка s\_dmacrm. bureaureqloan\_sstat |
| crd\_hist\_dmacrm\_clientcalls\_stran | Загрузка s\_dmacrm. clientcalls\_stran |
| crd\_hist\_dmpp\_clientbalinfo\_sagg |  |
| crd\_hist\_dmpp\_clientprofile\_sstat |  |
| crd\_hist\_dmpp\_dealstate\_invest\_sstat |  |
| crd\_hist\_dmpp\_dealstate\_sstat |  |
| crd\_hist\_dmpp\_profitabilitystate\_sstat |  |
| crd\_hist\_dwh\_transaction\_htran |  |
| crd\_hist\_dws132\_ad\_assets132\_mirror |  |
| crd\_hist\_ods\_whbranch\_oper |  |

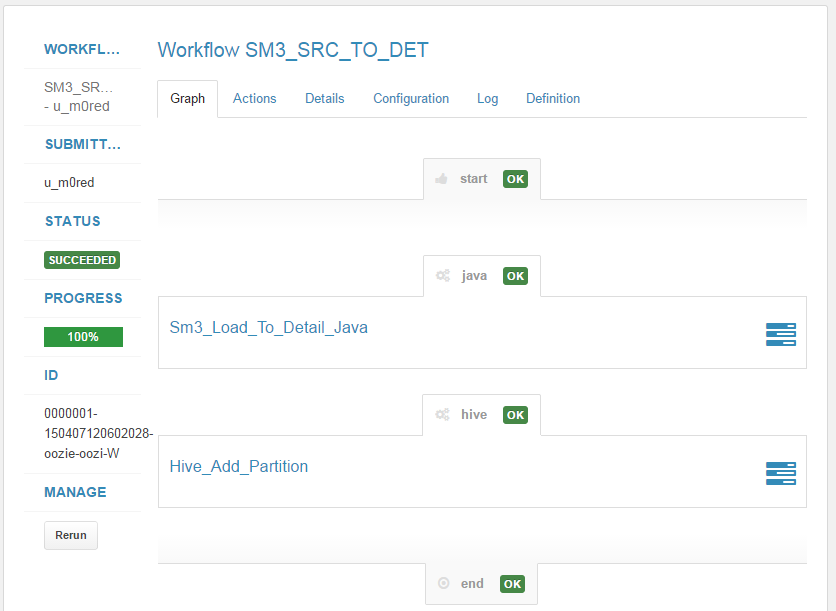
# Мониторинг работы системы

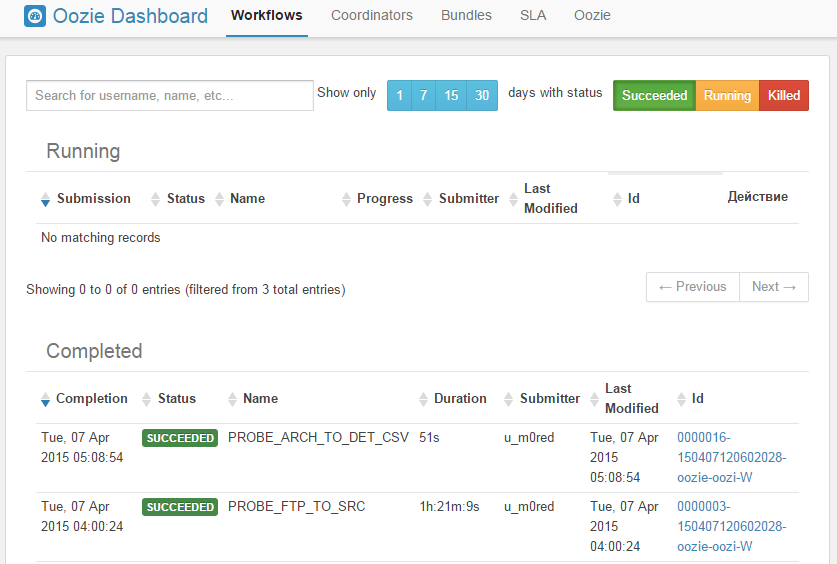
## Мониторинг журнала работы workflow

После запуска workflow можно наблюдать их работу, выбрав в HUE Web IU меню «Workflow» -> «Dashboard».



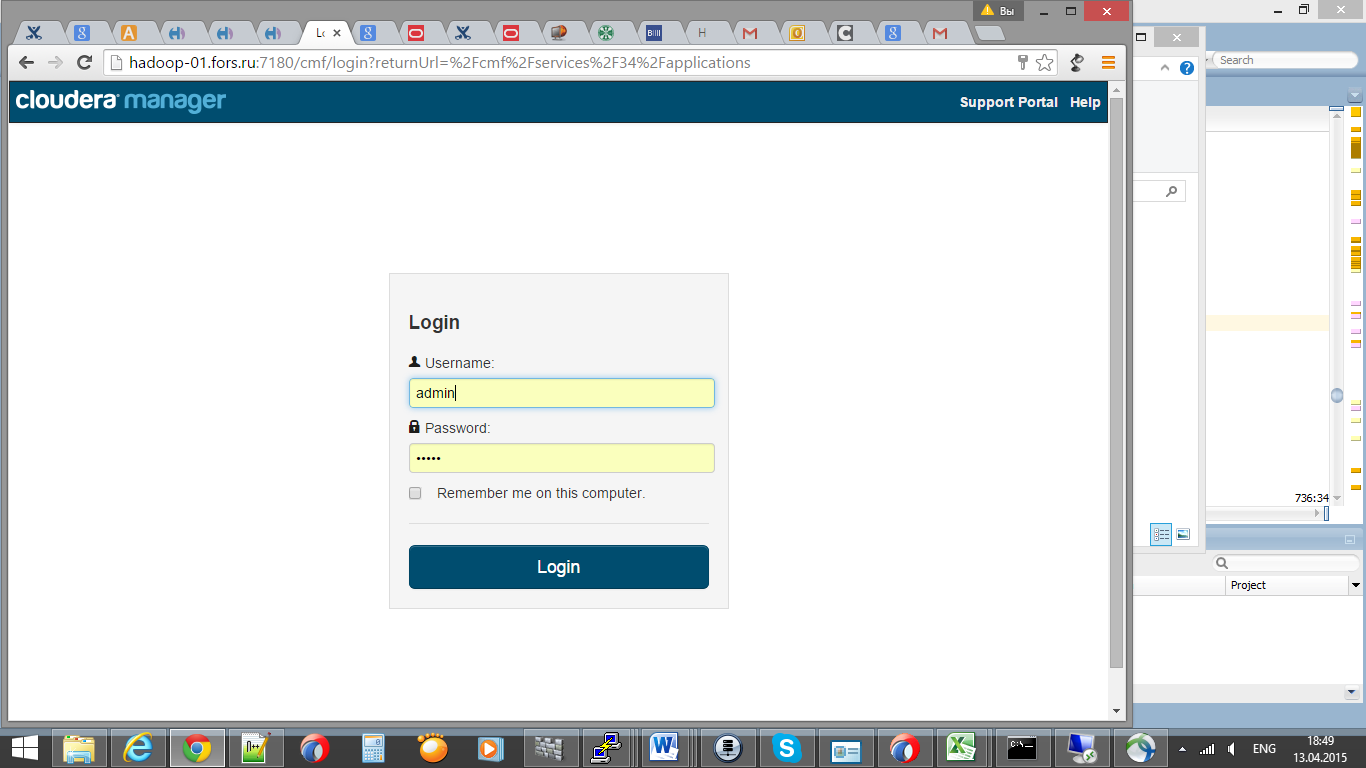
После успешного завершения работы загрузки её статус изменится на зелёный «SUCCEEDED», в случае завершения с ошибкой статус будет установлен как красный «FAILED».



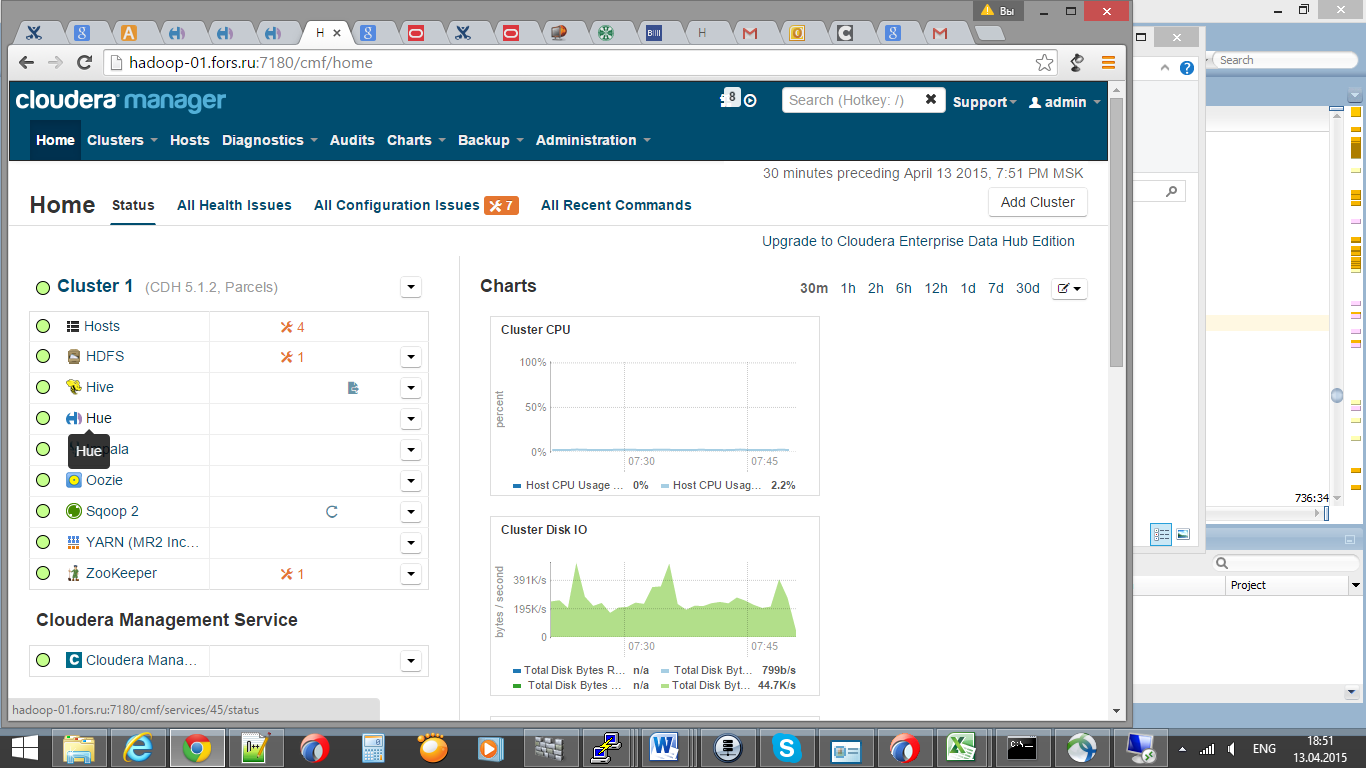


Для настройки количества отображаемых строк в журнале Oozie Dashboard необходимо настроить переменную «oozie\_jobs\_count». Для этого выполните следующие шаги (выполняется пользователем с ролью «Администратор»).

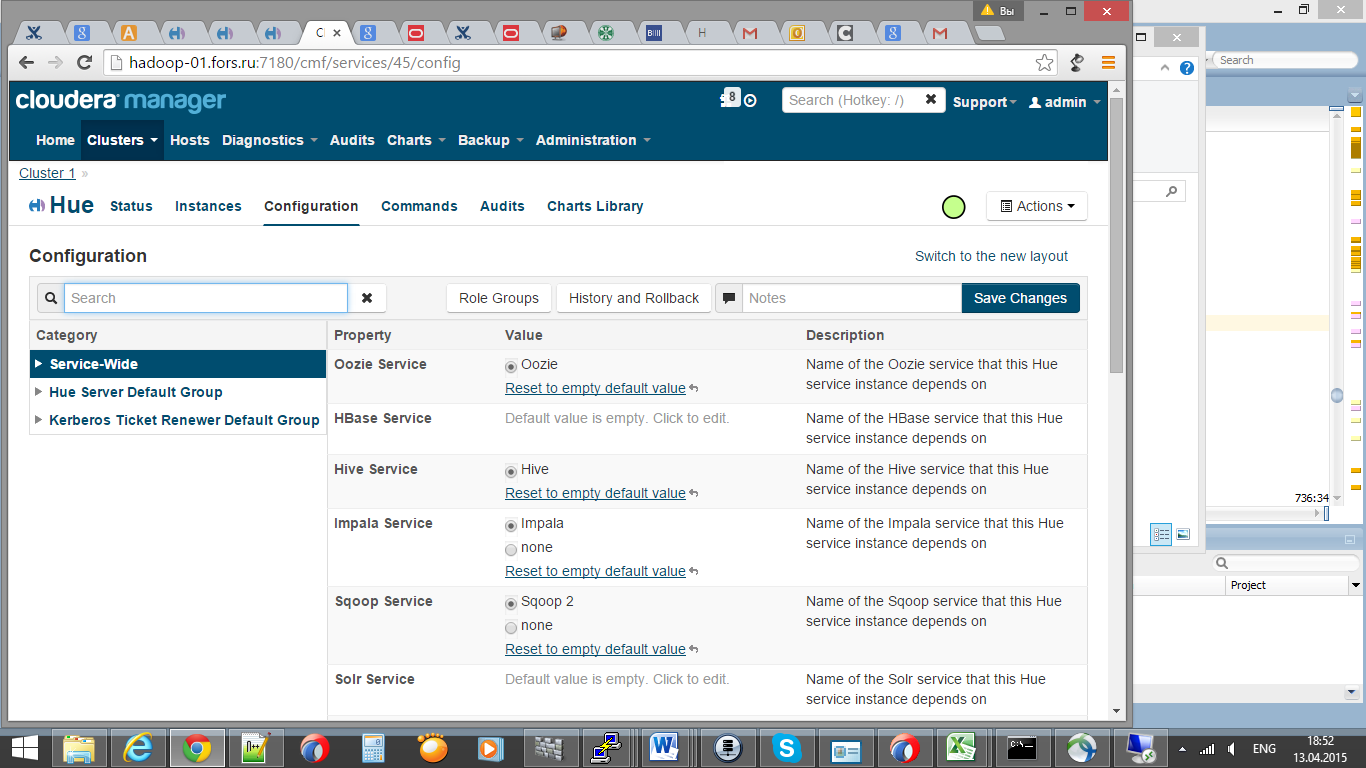
1. Зайти в интерфейс Cloudera Manager под администратором (<http://cm-host:7180>):



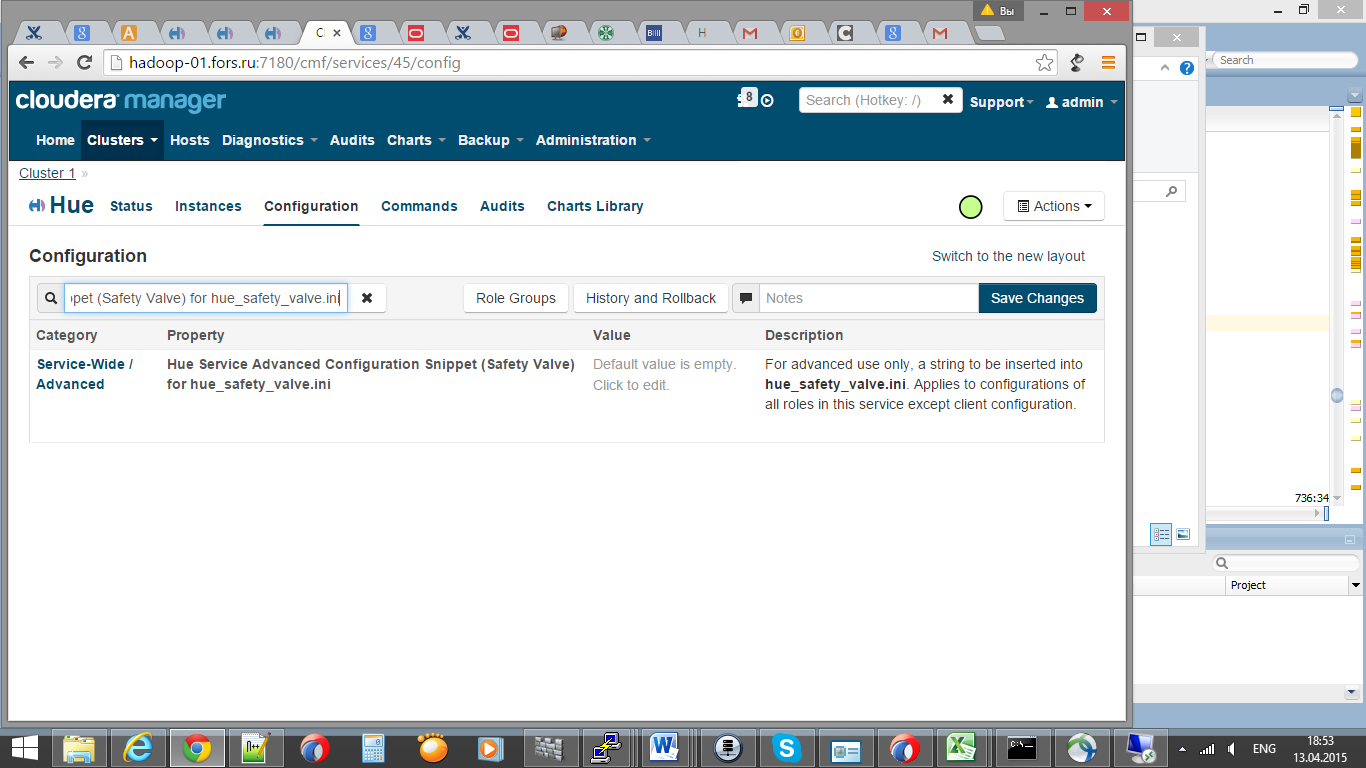
1. В списке сервисов выбрать сервис Hue:



1. Перейти на закладку «Configuration»:



1. В строке поиска ввести «Hue Service Advanced Configuration Snippet (Safety Valve) for hue\_safety\_valve.ini»:

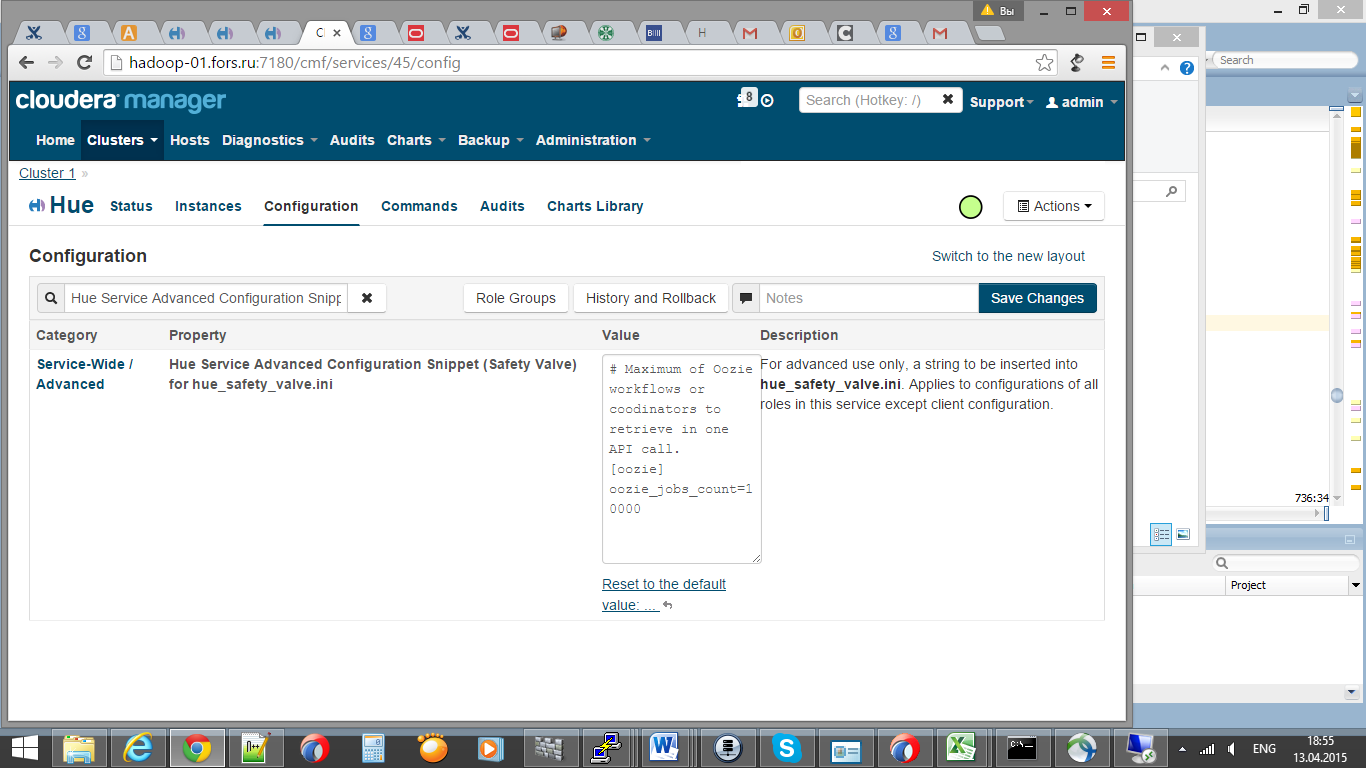


1. Добавить следующие строки в значение для найденного параметра:

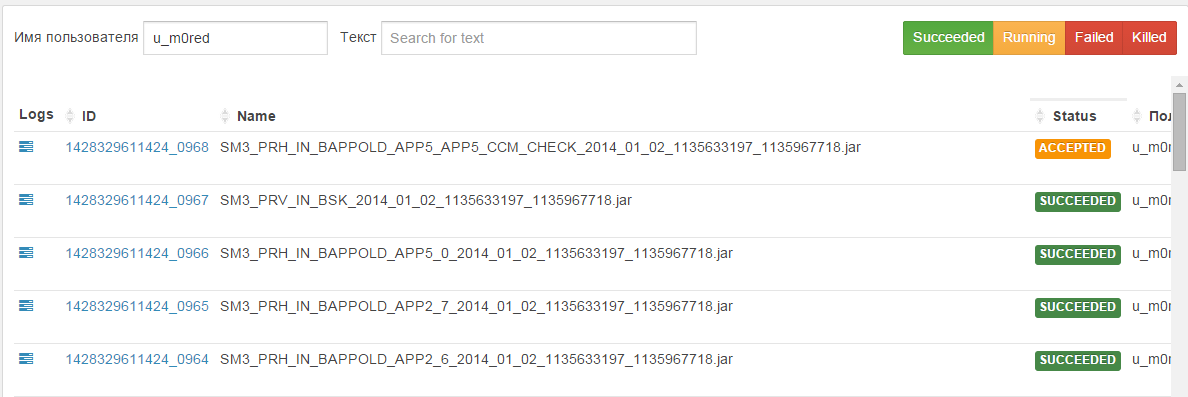
# Maximum of Oozie workflows or coodinators to retrieve in one API call.

[oozie]

oozie\_jobs\_count=10000



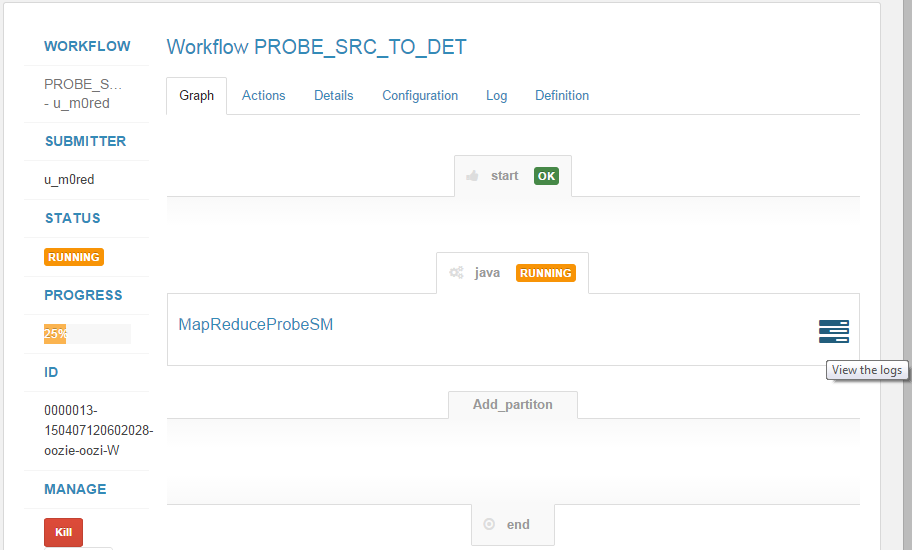
Oozie Dashboard отображает выполнение только задач системы, запущенных как workflow или coordinator. Все внутренние Map-Reduce задания, которые запускаются из workflow, здесь не видны. Для просмотра этих заданий можно воспользоваться HUE Job Browser:



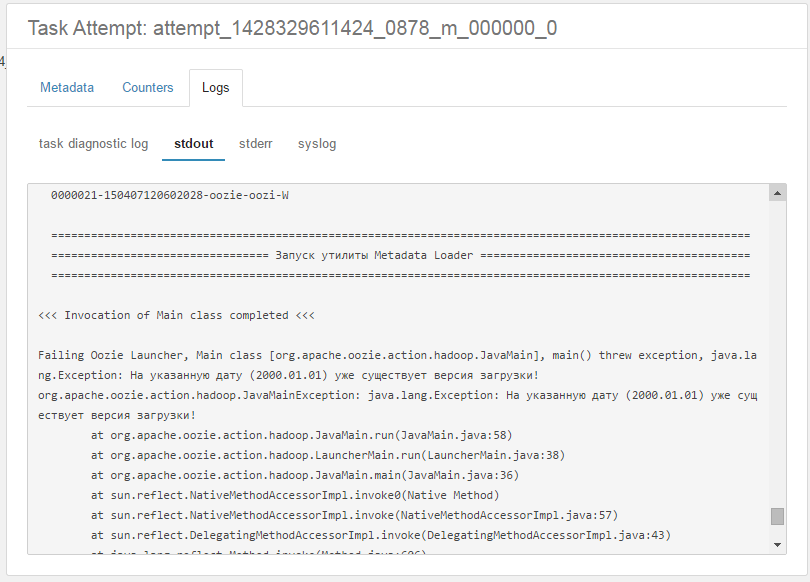
# Логирование

Система выполняет логирование в журнал выполнения Oozie workflow (полный журнал работы, включая отладочную информацию и ошибки) и в БД Метаданных (основные шаги работы workflow), для чего используется таблица META.RUN\_LOG. SQL-скрипты, которые формируются и выполняются в процессе загрузок отдельно сохраняются в каталог <hdfs\_home>/script/ (скрипты накапливаются за всю историю работы, однотипные скрипты разделены суффиксом «гггг.ММ.дд чч.мм.сс», соответствующему времени создания файла).

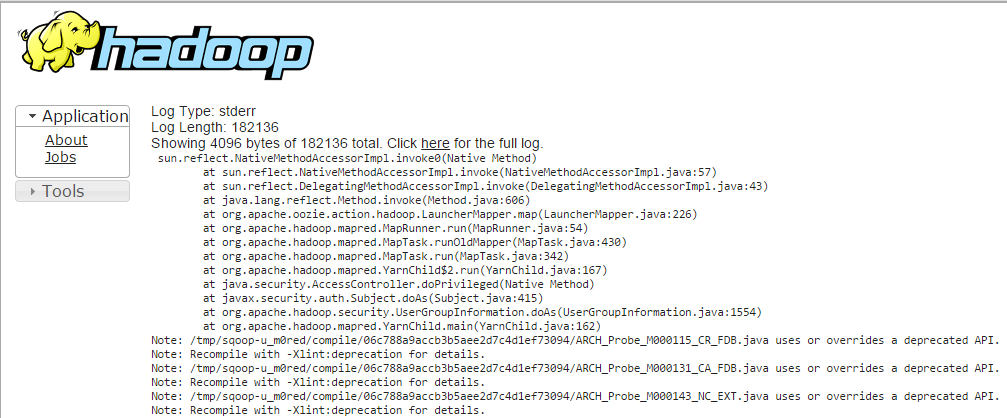
Журнал (лог) работы Oozie workflow можно посмотреть, перейдя на форму запущенного workflow и нажав ссылку «View the logs» (внутри блока action):



Стандартный выходной поток сообщений журнала отображается на вкладке «stdout».



Отображаемый журнал выводится из журналов YARN Applications, которые запускает Oozie. Как альтернатива, аналогичный журнал можно наблюдать в Web GUI Hadoop JobTracker (URL по умолчанию <http://yarn-host:8088/cluster>).



Помимо системных журналов HUE и Hadoop, так же существует журнал приложения (по умолчанию, время хранения информации – 3 года). Журнал хранится в БД Метаданных, в таблице META.RUN\_LOG, каждая запись привязана к Oozie Job ID (значение отображается в Oozie Dashboard и на форме запущенной workflow).

Файлы скриптов в каталоге <hdfs\_home>/script/ имеют следующие наименования: