Hadoop and Spark User Forum

Emil Kleszcz, Zbigniew Baranowski
IT-DB-SAS

Interacting with Hadoop clusters

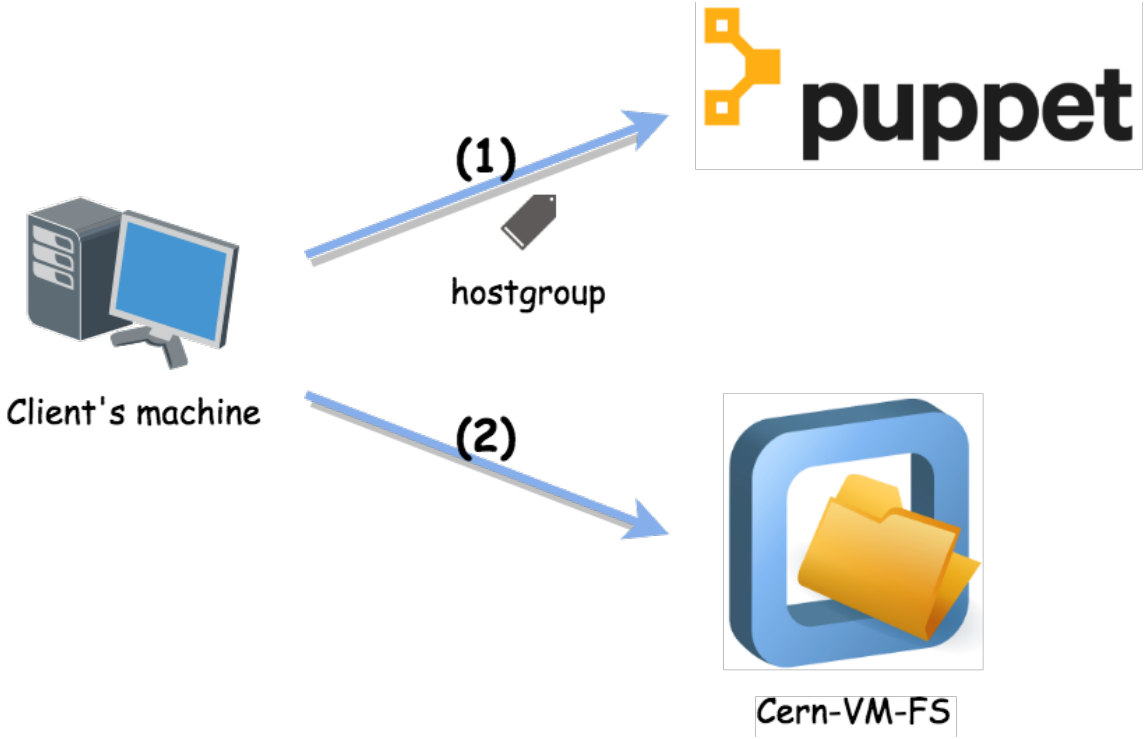
- What is needed to use the Hadoop and Spark clusters?
  - config files, binaries
  - e-group-based ACL granted, valid Kerberos ticket

- Supported ways of accessing the service (today)
  - 1) ~~Ssh to a cluster machine~~ (will be removed as of the end of August)
  - 2) Puppet client module
  - 3) Sourcing environments from CVMFS
    - LXPLUS or any machine with CVMFS (see KB0004426)
  - 4) Using SWAN for interactive analysis **(PYSPARK)**

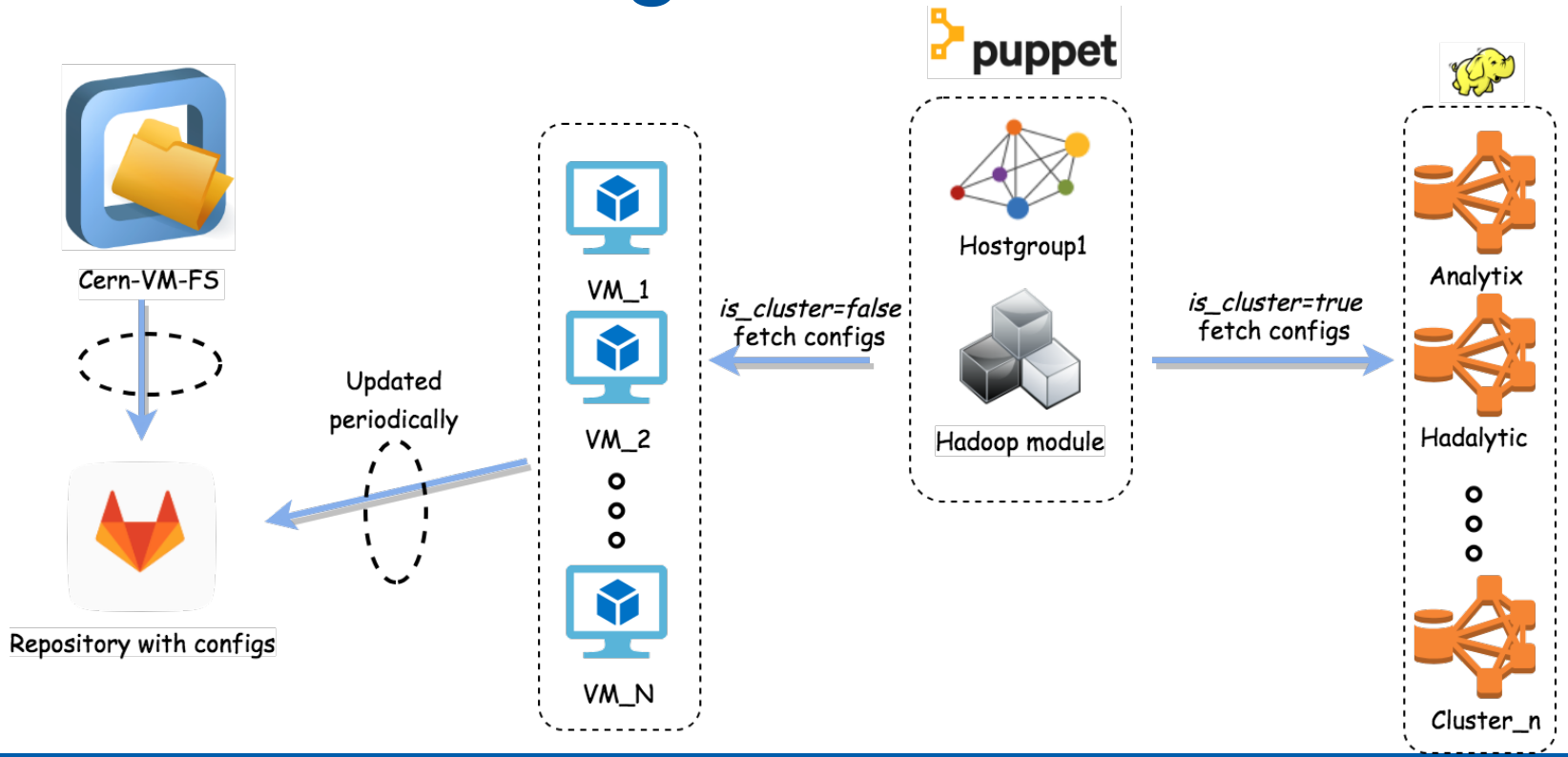
Limitations of the current setup

- Two configuration workflows to maintain
  - Puppet
  - CVMFS
- Automatically updated configuration possible only with puppet
  - There are communities that do not use puppet
- Puppet module is suboptimal
  - Single module for server and client
  - Executes all the server logic on a client machine
  - Takes time to execute
  - Configuration files contain to much information
- Extraction of service configuration for CVMFS is complex
  - Virtual machine per cluster to fetch and update new configuration

Current ways for config extraction



Current config extraction



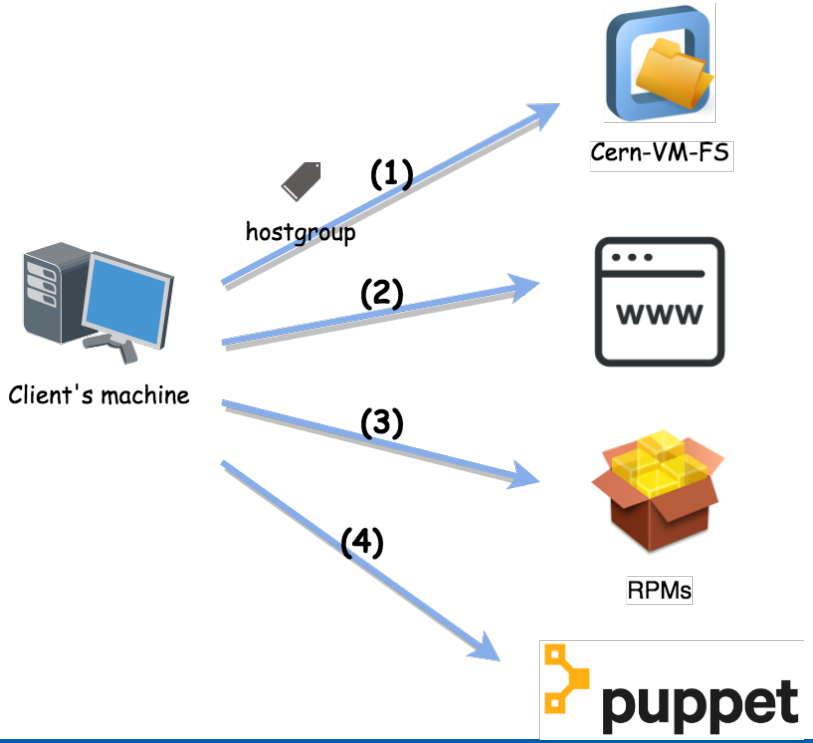
New Hadoop Client

- **Simplicity** – clean set of required config files, unify the configuration workflow for all client use cases
- **Performance** – simple puppet module separated from the server module
- **Automation** – automatic updates (cron job), no VM per cluster needed
- **Accessibility** – various ways to configure, depending on needs
- **Flexibility** – one can easily connect to another cluster, without major reconfiguration

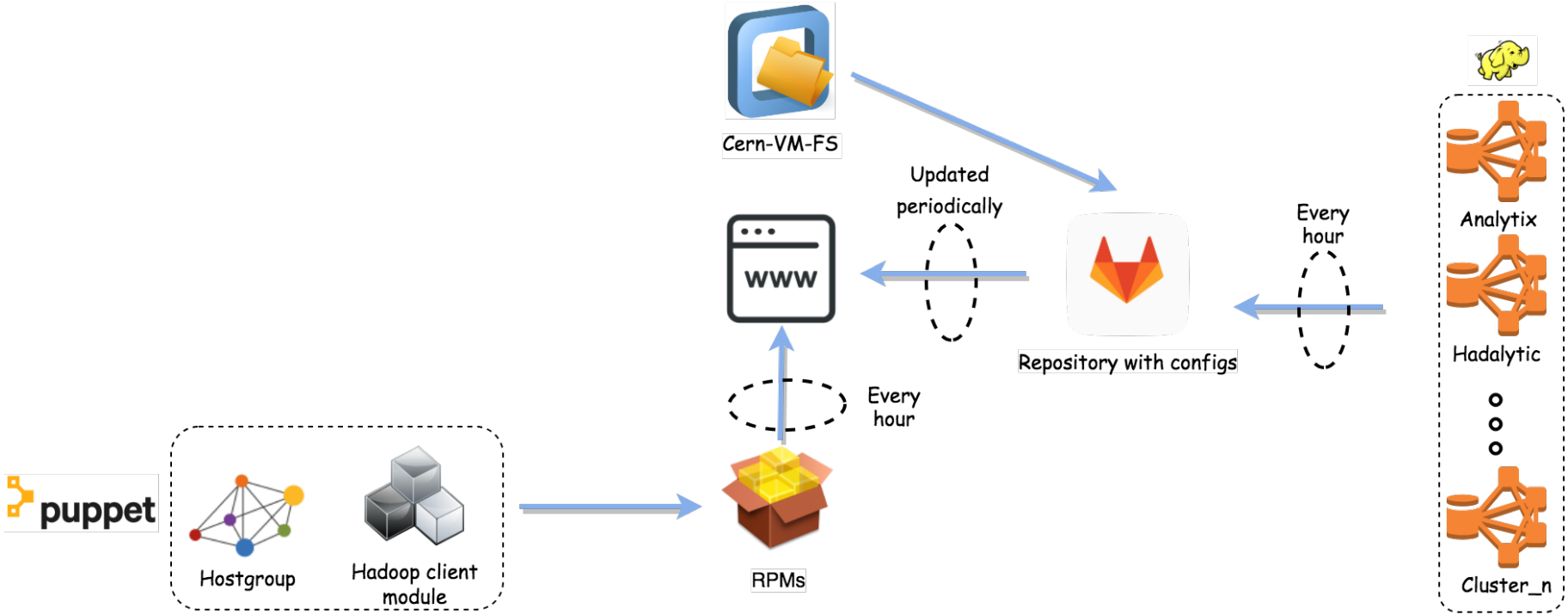
New Hadoop configuration - client

- Config files accessible via:
  - Web server (https://hadoop-config.web.cern.ch/files/hadoop/)
  - RPMs
  - Puppet module (https://gitlab.cern.ch/ai/it-puppet-module-hadoop/tree/new_client)
  - CVMFS
- Separated from server components
- Clean structure of files and scripts

New ways for config extraction



New config extraction



Files structure on www

- clusters.txt <- list of available clusters
- conf/ <- configuration for each component (incl. Hadoop, HBase, Hive, and Spark)
- hadoop-configs.tar.gz <- all scripts and conf.

Setting up target cluster for client

- Configuration per cluster & per Hadoop component
- Script for default configuration – run once

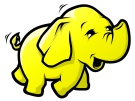
```
hadoop-default-set-conf.sh cluster_name
```

- configuration per shell session

```
source hadoop-setconf.sh cluster_name
```

RPM packages

- cern-hadoop-client
  - cern-hadoop-config
  - spark-bin-2.3.0
  - hadoop-bin-2.7.5
  - hbase-bin-1.2.6
  - java-1.8.0-openjdk
- cern-hadoop-xrootd-connector
  - For EOS and Root files use case
- Available in QA repository (hdp7-qa)
  - http://linuxsoft.cern.ch/internal/repos/hdp7-qa/x86_64/os/Packages/



Other dependencies

- Deployed
  - Java Development Kit ($_{JAVA\_HOME}$) – installed automatically with the RPMs
- Required
  - Kerberos valid ticket (*kinit*)
  - For Spark – enabled ports for incoming traffic on a block manager and a driver
  - Sufficient permissions for scripts execution
  - Yellowdog Updater, Modified (YUM)

Installation with Puppet

1. Include new Hadoop module in your hostgroup manifest (new_client branch)

```
include ::hadoop
```

2. For accessing different cluster than analytix parameterize the module with a cluster name via Hiera *parameter*

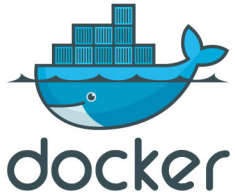
```
# In data/hostgroup/<hostgroup_name>/<class_name>.yaml
hadoop::cluster_name: hadalytic
```

3. Supporting backward compatibility for current way of including the module –> no need to change manifests

More information in the User guide

- Client configuration step-by-step
- Troubleshooting
- User guide:
  - https://hadoop-user-guide.web.cern.ch/hadoop-user-guide
  - KB0005494

Future work and ideas



- Docker images integrated into GitLab
  - With Docker Container Registry
  - Support for other OSs

- Available at the beginning of July

Timeline

- Available for testing since 1st of June
  - module: hadoop , branch: new_client
- 11th of June available on QA branch
- 28th of June on master branch
- 2nd of July - Docker images
- 28th of August full transition to the new Hadoop module (no backward compatibility support)
- 28th of August access via ssh revoked

Thank you for your attention