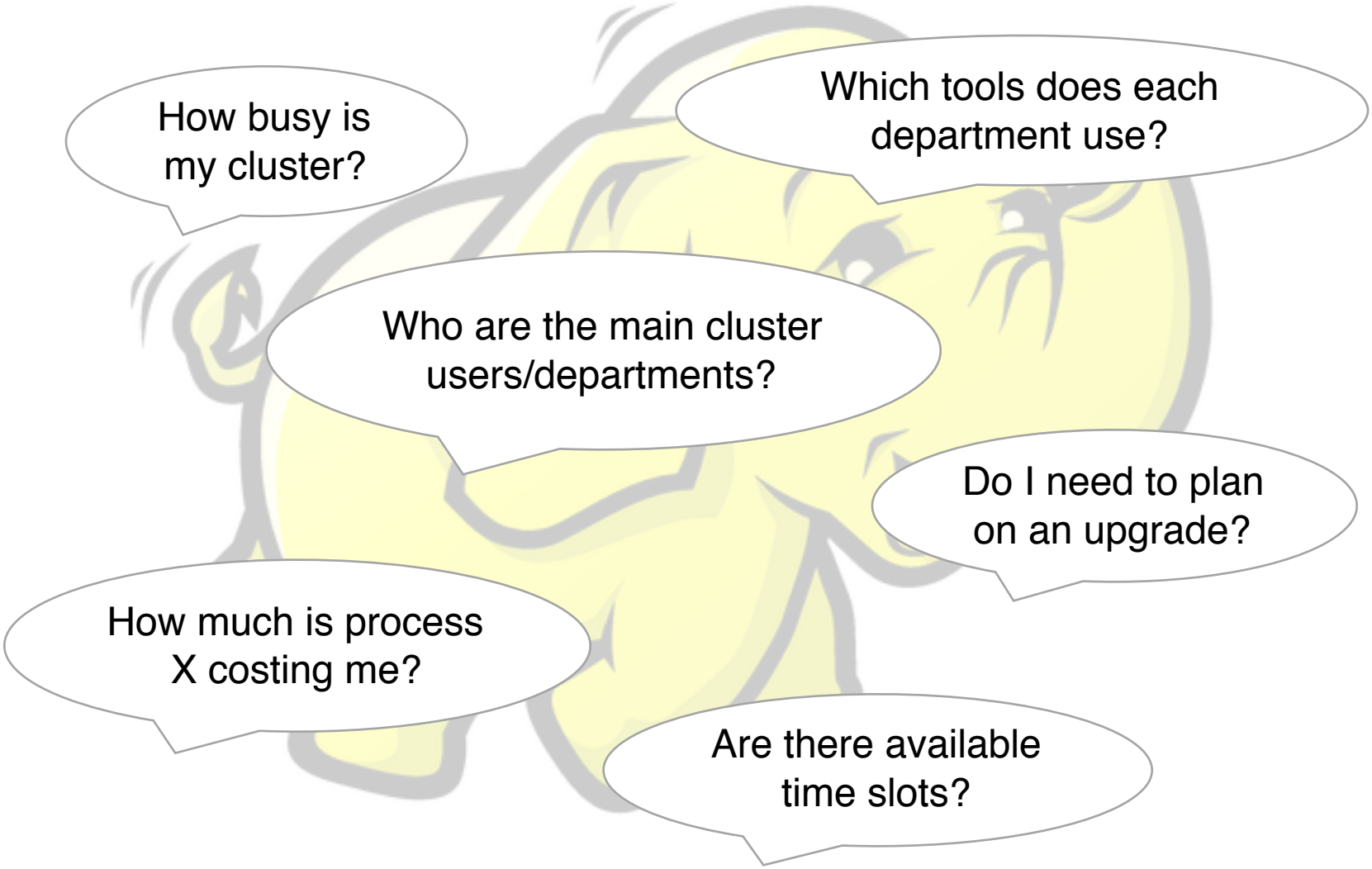
TAMING BIG DATA

André Simões
andre.simoes@xpand-it.com
@ITXpander

xpandit

GENERAL ISSUES



(SOME OF) THE PROBLEMS

ORCHESTRATION IS **HARD**

MONITORING IS **HARDER**

NOTIFICATIONS… WELL...

ONE WAY TO DO IT...

CONTROLLING THE ETL LAUNCH

CENTRALIZED ORCHESTRATION

POOL MANAGEMENT

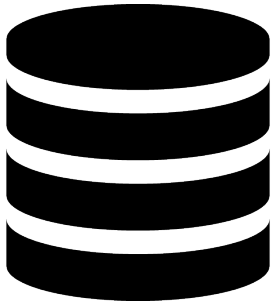
EXECUTION DELEGATION

ORCHESTRATION METADATA

REMOTE AGENTS

CLUSTER DATA

DEPENDENCIES



SCHEDULING

ETL METADATA

POOL METADATA

USER MAPPING

REMOTE AGENTS

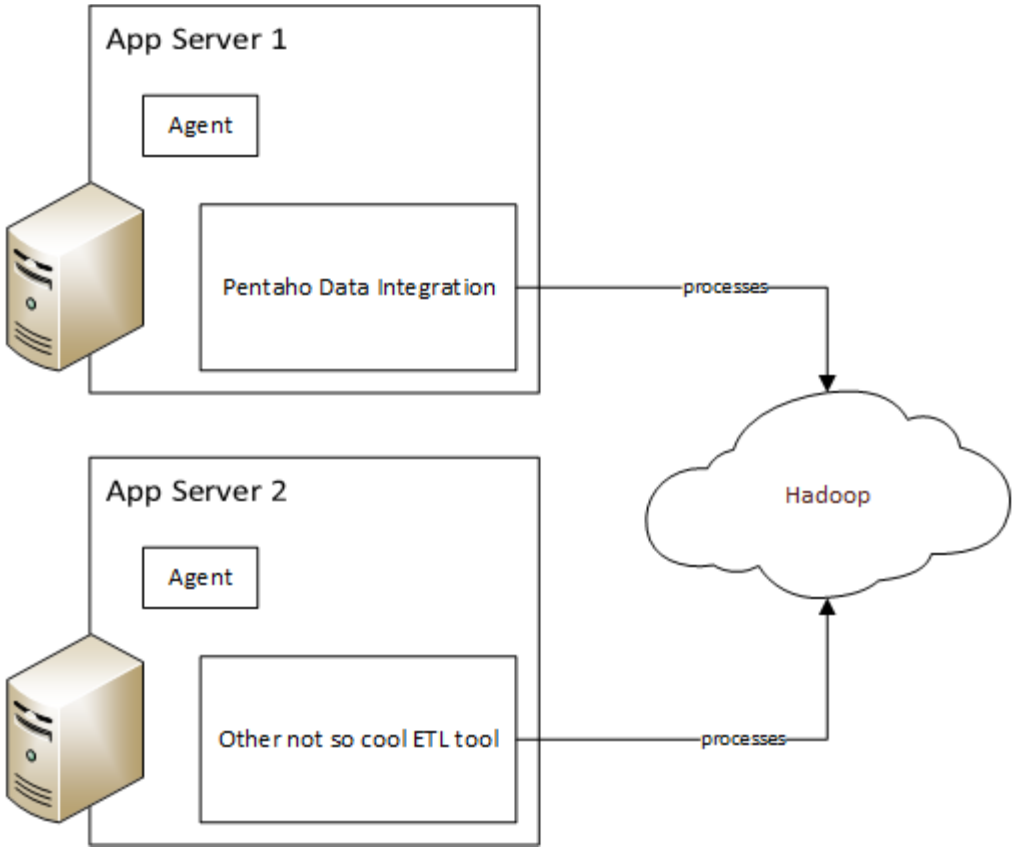
FINE GRAINED CONTROL

ETL TOOL SPECIFIC

REAL TIME LOGGING

ERROR RECOVERY

ASYNC EXECUTION



GATHERING EXECUTION DATA

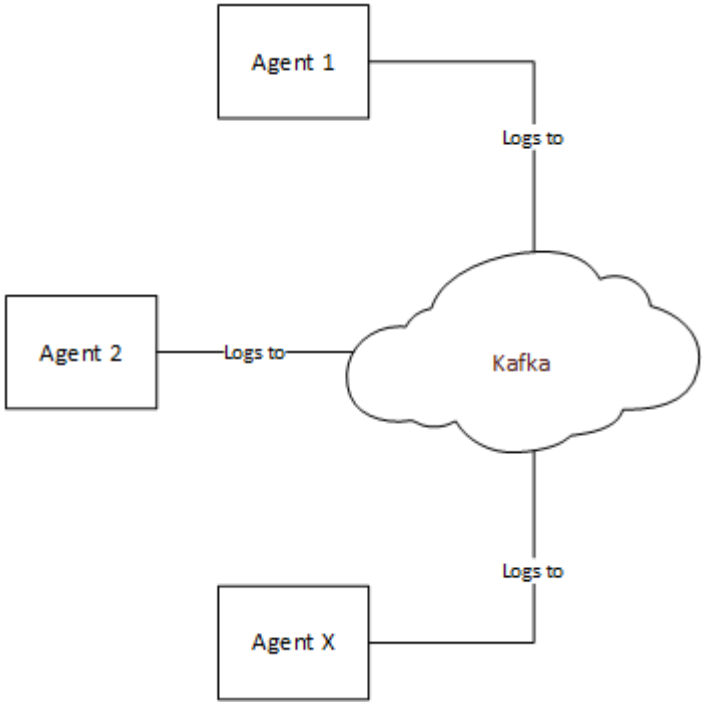
USE KAFKA AS A LOG SINK

FAULT TOLERANT

REAL TIME

CONSISTENT

CONTROLLED BY THE AGENT



PDI EXTENSION POINTS

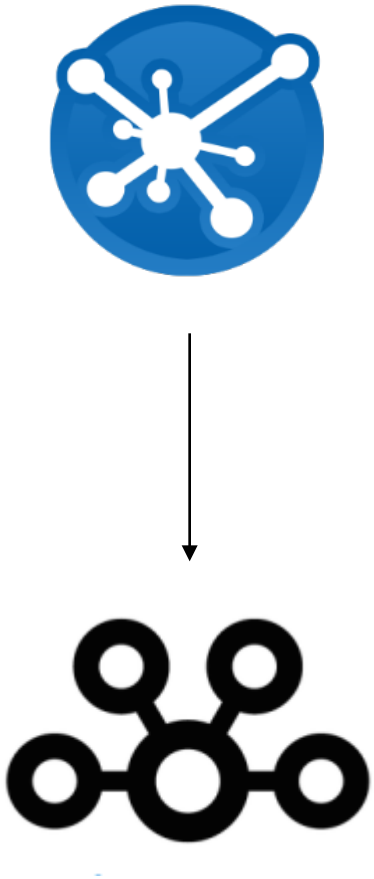
CAPTURE LOG START

CAPTURE LOG END

CAPTURE CONNECTION TYPE

CAPTURE STEP LINEAGE DETAIL

GENERATES NOTIFICATIONS



COLLECT LOG DATA IN (AS) REALTIME (AS POSSIBLE)

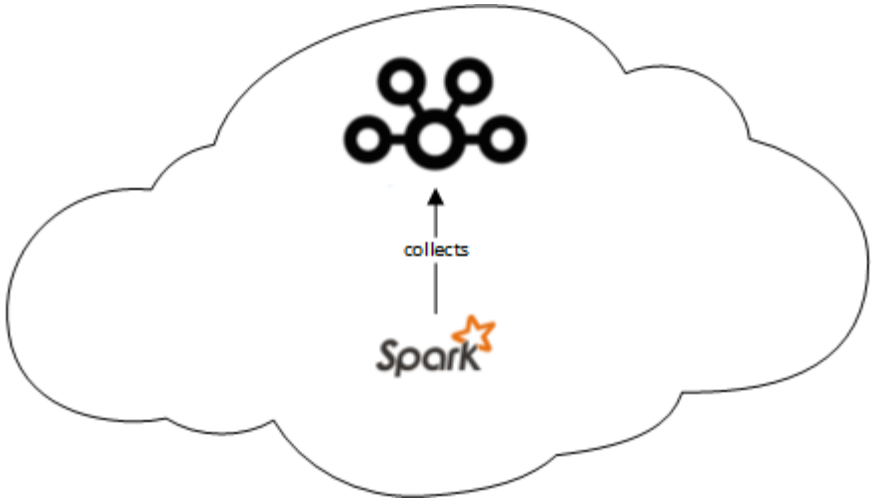
SPARK AS KAFKA COLLECTOR

REAL TIME LOG PARSING

ETL TOOL ADAPTABLE

DATA DUMPS IN IMPALA AND HBASE

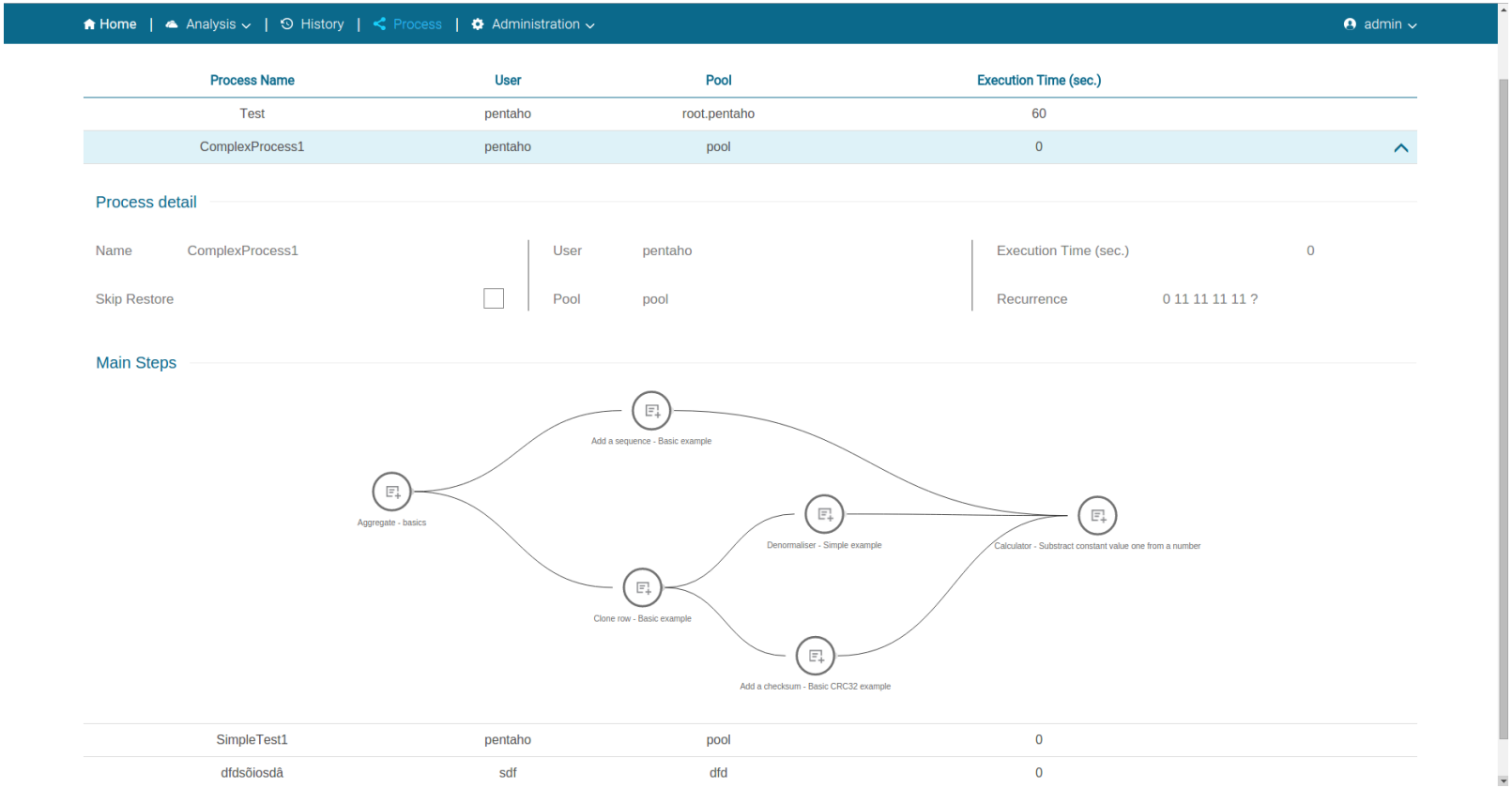
GENERATES NOTIFICATIONS



BUT...

... HOW DOES IT LOOK LIKE?

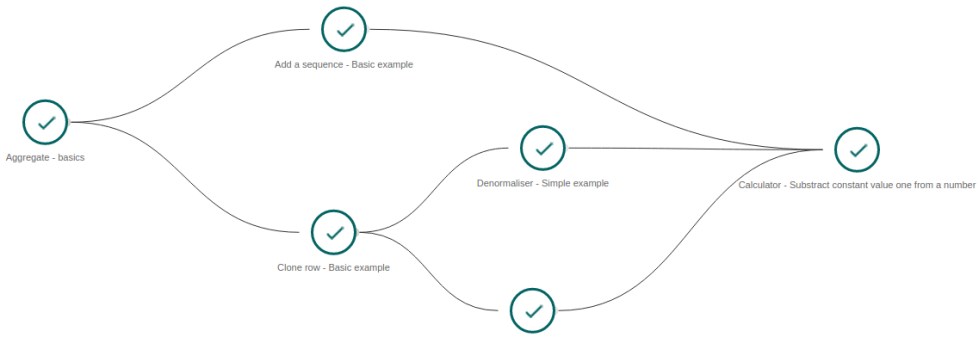
CONFIGURATION



EXECUTION HISTORY

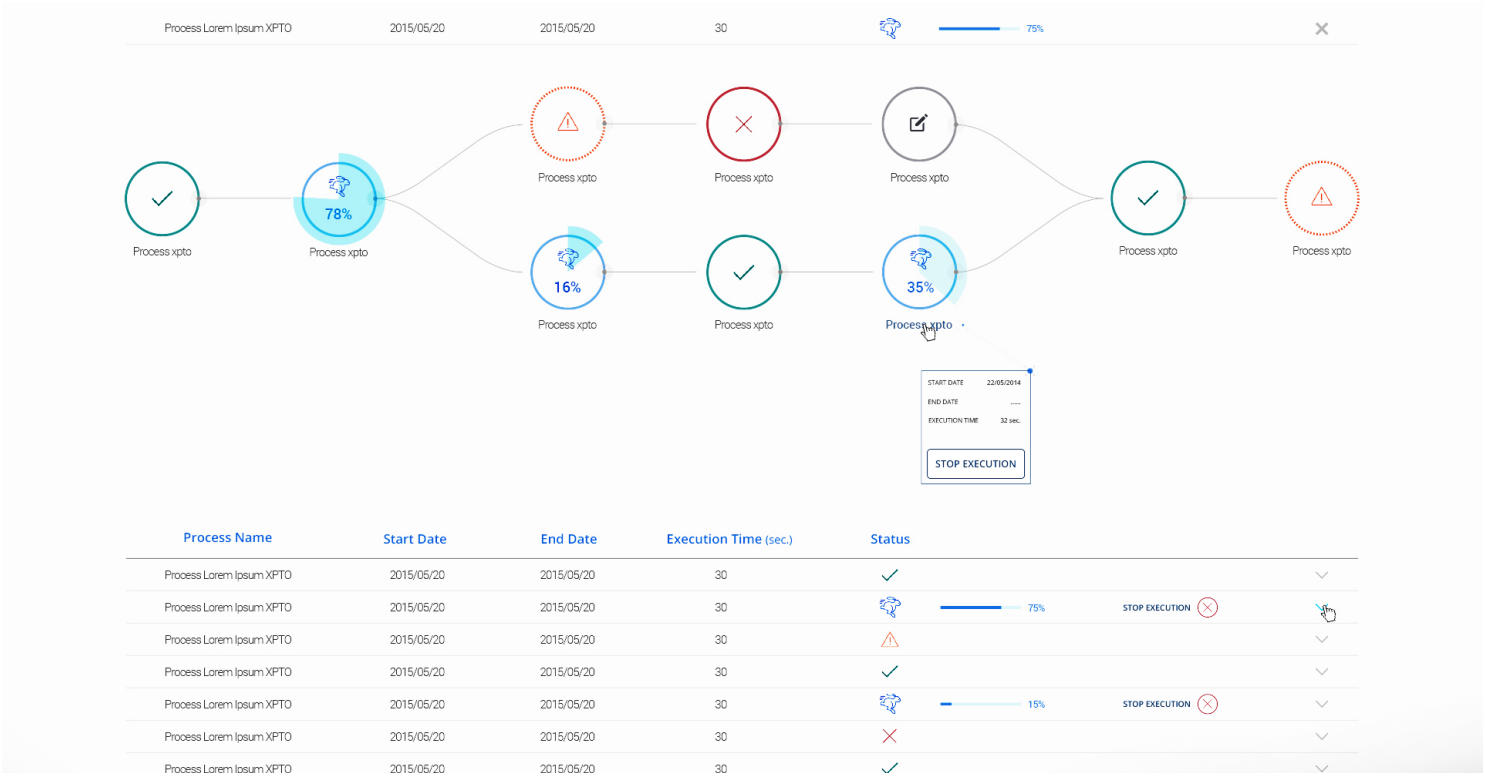
Home | Analysis | History | Process | Administration — admin

Process Name	Start Date	End Date	Status	Execution time (sec.)
SimpleTest1	2015-09-09 15:28:28.882	2015-09-09 15:28:37.205	Finished	8
SimpleTest1	2015-09-09 15:08:25.113	2015-09-09 15:08:36.908	Finished	11
SimpleTest1	2015-09-09 13:50:06.085	2015-09-09 13:50:14.086	Finished	8
SimpleTest1	2015-09-09 12:59:00.454	2015-09-09 12:59:08.413	Finished	7
SimpleTest1	2015-09-09 12:58:29.894	2015-09-09 12:58:37.319	Finished	7
SimpleTest1	2015-09-09 12:55:50.16	2015-09-09 12:56:35.613	Finished	45
ComplexProcess1	2015-11-02 18:13:56.047	2015-11-02 18:13:58.7	Crashed	2
ComplexProcess1	2015-09-23 12:12:24.662	2015-09-23 12:12:29.087	Crashed	4
ComplexProcess1	2015-09-09 16:41:07.014	2015-09-09 16:41:38.338	Finished	31



ComplexProcess1	2015-09-09 15:26:10.042	2015-09-09 15:26:49.108	Finished	39
ComplexProcess1	2015-09-09 15:09:59.71	2015-09-09 15:10:34.311	Finished	34

REALTIME (AS POSSIBLE) DATA



ANALYTICS

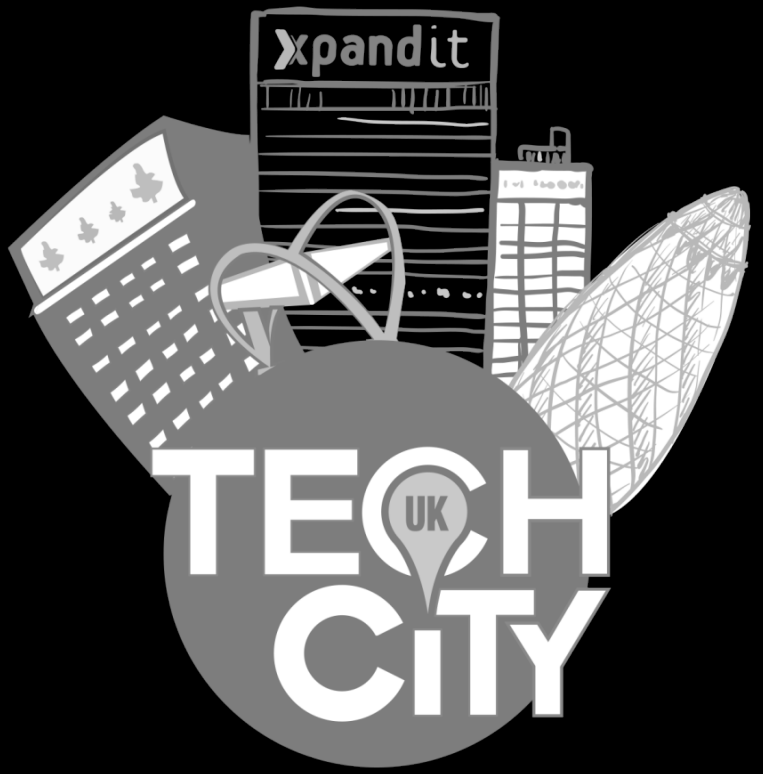
Intentionally left blank.

THE BAD AND THE UGLY

LIMITED POOL MANAGEMENT

MANAGING KERBERIZED CLUSTERS

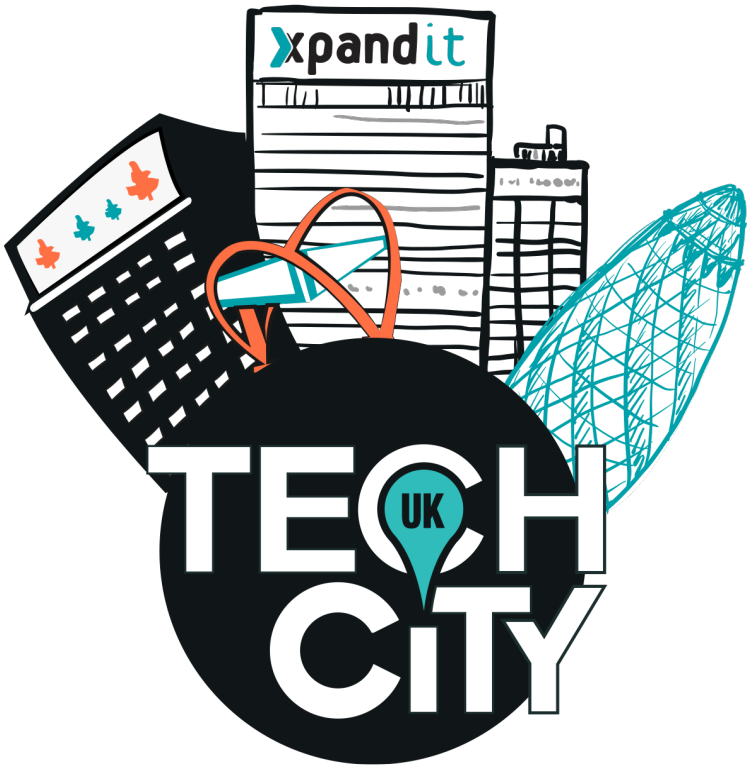
MULTI-POINT DEPLOYMENT



NEW LONDON OFFICE!

THANK YOU!

NEWSFLASH



We are now in London!