### Hello User, Welcome to Database

Joe DeVivo works for Basho, tweets at @joedevivo, and put some relevant files at <u>github.com/joedevivo/ricon</u> If people can't set up your product, they secretly hate you

I found an amazing database back in 2010.

It wasn't too hard to configure back then

turns out, that was the only part that didn't scale

```
%% -*- tab-width: 4;erlang-indent-level: 4;indent-tabs-mode: nil -*-
%% ex: ts=4 sw=4 et
%% Riak Core config
 {riak core, [
   %% Default location of ringstate
    {ring_state_dir, "data/ring"},
   %% riak_web_ip is the IP address that Riak's HTTP interface will
   %% bind to. If this is undefined, the HTTP interface will not run.
    {web ip, "127.0.0.1"},
   %% riak web port is the TCP port that Riak's HTTP interface will
   %% bind to.
     {web_port, 8098}
    ]},
%% Riak KV config
 {riak kv, [
   %% Storage_backend specifies the Erlang module defining the storage
   %% mechanism that will be used on this node.
    {storage_backend, riak_kv_dets_backend},
   %% Different storage backends can use other configuration variables.
   %% For instance, riak_dets_backend_root determines the directory
   %% under which dets files will be placed.
    Iniak ky dots backond noot "data/dots"l
```

```
%% -*- tab-width: 4;erlang-indent-level: 4;indent-tabs-mode: nil -*-
%% ex: ts=4 sw=4 et
[
%% Riak Core config
 {riak_core, [
              %% Default location of ringstate
              {ring_state_dir, "data/ring"},
              %% riak_web_ip is the IP address that Riak's HTTP interface will
              %% bind to. If this is undefined, the HTTP interface will not run.
              {web_ip, "127.0.0.1"},
              %% riak_web_port is the TCP port that Riak's HTTP interface will
              %% bind to.
              {web_port, 8098}
             ]},
 %% Riak KV config
 %% mechanism that will be used on this node.
            {storage_backend, riak_kv_dets_backend},
            %% Different storage backends can use other configuration variables.
            %% For instance, riak_dets_backend_root determines the directory
            %% under which dets files will be placed.
            {riak_kv_dets_backend_root, "data/dets"},
            %% riak_handoff_port is the TCP port that Riak uses for
            %% intra-cluster data handoff.
            {handoff_port, 8099},
            %% pb_ip is the IP address that Riak's Protocol Buffers interface
            %% will bid to. If this is undefined, the interface will not run.
            {pb_ip, "0.0.0.0"},
            %% pb_port is the TCP port that Riak's Protocol Buffers interface
            %% will bind to
            {pb_port, 8087},
            %% raw_name is the first part of all URLS used by Riak's raw HTTP
            %% interface. See riak_web.erl and raw_http_resource.erl for
            %% details.
            %{raw_name, "riak"},
            %% mapred_name is URL used to submit map/reduce requests to Riak.
            {mapred_name, "mapred"},
            %% js_vm_count is the number of Javascript VMs to start per Riak
            %% node. 8 is a good default for smaller installations. A larger
            %% number like 12 or 16 is appropriate for installations handling
            %% Lots of map/reduce processing.
            {js_vm_count, 8},
            %% js_source_dir should point to a directory containing Javascript
            %% source files which will be loaded by Riak when it initializes
            %% Javascript VMs.
            %{js_source_dir, "/tmp/js_source"}
            %% riak_stat enables the use of the "riak-admin status" command to
            %% retrieve information the Riak node for performance and debugging needs
            {riak_kv_stat, true}
           ]},
 %% SASL config
 {sasl, [
         {sasl_error_logger, {file, "log/sasl-error.log"}},
         {errlog_type, error},
         {error_logger_mf_dir, "log/sasl"},  % Log directory {error_logger_mf_maxbytes, 10485760},  % 10 MB max file size
         {error_logger_mf_maxfiles, 5}
                                                % 5 files max
         ]}
].
```

```
## Name of the riak node
-name dev1@127.0.0.1
## Cookie for distributed erlang
-setcookie riak
## Heartbeat management; auto-restarts VM if it dies or becomes unresponsive
## (Disabled by default..use with caution!)
##-heart
## Enable kernel poll and a few async threads
+K true
+A 5
## Increase number of concurrent ports/sockets
-env ERL MAX PORTS 4096
## Tweak GC to run more often
-env ERL FULLSWEEP AFTER 10
```

Two configuration files with different syntaxes!

That's sure to win us database of the year!

# Guess how many times I broke this? A lot

#### Forget a comma? Your configuration broke

# Didn't close that bracket? Too bad

# Not an Erlang developer? Well, you should be

### Erlang





### LOL, Syntax

## Erlang developers write Erlang friendly configuration settings

```
• {on, []}
```

```
• {off, []}
```

{on, [debug]}

#### one setting, lots of lines

```
actual number!
```

```
W raw_mame is the first part of all URLS used by the Risk raw
A interface. See risk_web_eri and raw_http_resource_eri for
A details.
applications the state of the s
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  hyperbole!
```

warning: contains about a billion
[]'s, {}'s, commas, and periods to misplace

We addin pasel. {userlist, {{"user", "pass"} }},

1), Wh The admin panel is broken up into multiple Wh components, each of which is enabled or disabled Wh by one of these settings. (admin, true)

#### NOTHING!

Everything you need to know about a single configuration setting should be on a single line

That line can be anywhere in the file

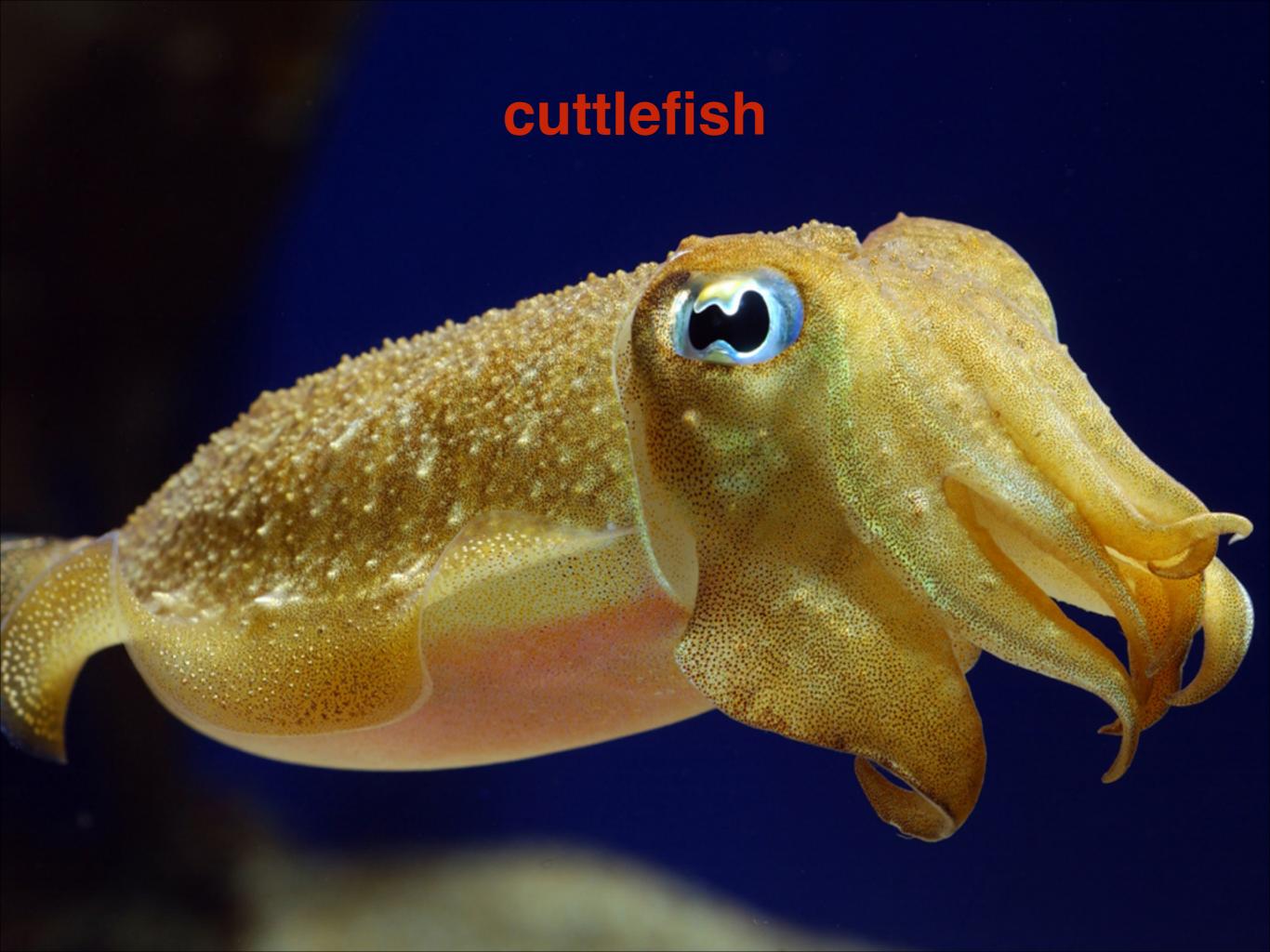
# If you set something more than once, the last one wins



Key/Value is a pretty good idea

What did I want to give the user?

```
## enable active anti-entropy subsystem
## possible values: on, off, debug
anti_entropy = on
```



```
## Kestrict now fast AAE can bulla hash trees. Bullaing the tree
## for a given partition requires a full scan over that partition's
## data. Once built, trees stay built until they are expired.
## Config is of the form:
## {num-builds, per-timespan}
## Default is 1 build per hour.
anti_entropy.build_limit.number = 1
anti_entropy.build_limit.per_timespan = 1h
## Determine how often hash trees are expired after being built.
## Periodically expiring a hash tree ensures the on-disk hash tree
## data stays consistent with the actual k/v backend data. It also
## helps Riak identify silent disk failures and bit rot. However,
## expiration is not needed for normal AAE operation and should be
## infrequent for performance reasons. The time is specified in
## milliseconds. The default is 1 week.
anti_entropy.expire = 1w
## Limit how many AAE exchanges/builds can happen concurrently.
anti_entropy.concurrency = 2
## The tick determines how often the AAE manager looks for work
## to do (building/expiring trees, triggering exchanges, etc).
## The default is every 15 seconds. Lowering this value will
## speedup the rate that all replicas are synced across the cluster.
## Increasing the value is not recommended.
anti entropy.tick = 15s
```

```
→ rel git:(develop) / cat riak.conf | grep anti_entropy anti_entropy = on anti_entropy.build_limit.number = 1 anti_entropy.build_limit.per_timespan = 1h anti_entropy.expire = 1w anti_entropy.concurrency = 2 anti_entropy.tick = 15s anti_entropy.data_dir = ./data/anti_entropy anti_entropy.write_buffer_size = 4MB anti_entropy.max_open_files = 20
```

```
## Default ring creation size. Make sure it is a power of 2,
## e.g. 16, 32, 64, 128, 256, 512 etc
ring_size = 42
```

```
→ rel git:(develop) X ./riak/bin/riak console

10:11:27.284 [info] Application lager started on node nonode@nohost

10:11:27.284 [info] Checking /Users/joe/dev/basho/riak/rel/riak/bin/../

etc/app.config exists... false

10:11:27.284 [info] Checking /Users/joe/dev/basho/riak/rel/riak/bin/../

etc/vm.args exists... false

10:11:27.284 [info] No app.config or vm.args detected in /Users/joe/dev/basho/riak/rel/riak/bin/../etc, activating cuttlefish

10:11:27.717 [info] Adding Defaults

10:11:27.723 [info] Applying Datatypes

10:11:27.751 [info] Validation

10:11:27.751 [error] ring_size invalid, not a power of 2 greater than 1

10:11:27.751 [error] Some validator failed, aborting
```

While is not be sent the service of the sent to the sent cluster of the sent to the sent t

## Name of the risk node

Ww -- mode: erlang:erlang-indent-level: 4;indent-tabs-mode: nil ---Www.ftmerland.ts=4 su=4 et When the recommendation was enyou had filler the content of the content to the content of the Vk Default ring creation size. Make sure it is a power of 2, Vk e.g. 16, 32, 64, 128, 256, 512 etc Vering creation size, 64), We http is a list of IP addresses and TCP ports that the Riak We HTTP interface will bind. (http. [ "127.0.0.1", 10018 } )}, JUTTS, ("UIJABLY, NEWES) ]),
whites is a list of EV address and TD ports that the Elak
My HTMS difference will lack
My HTMS difference will see that
My HTMS difference will be that
My HTM We risk handoff port in the TCP port that Risk uses for We intra-cluster data handoff. (handoff\_port, 18019 ), VM To encrypt risk core intra-cluster data handoff traffic, VM uncomment the following line and edit its path to an VM appropriate certific and keyfile. (This example uses a VM ingle file with both items concatenated together.) VMpandoff all options ([certifile, "They|eriserve.pem\*]), vymmoser\_ms\_ops\_aus\_ (tenting, representations\_ms)

% Dirac support

% Do not emable farace\_support unless your Erlang/SIP

% united is compiled to support Dirace. Dirace is

% available in RISBSE (supported by the Erlang/SIP)

% official source package] and in RISBSE via a custom

% source repository & branch.

(drace\_support, failes) (firmer\_usport, faile),
Whomith Cheeks
Whomith
Whomith Cheeks
Whomith Cheeks
Whomith Cheeks
Whomith Cheeks
Whom Wh Platform-specific installation
(platform\_bin\_dir, "./bin"),
(platform\_data\_dir, "./data"),
(platform\_itd\_dir, "./ett"),
(platform\_lib\_dir, "./lib"),
(platform\_lib\_dir, "./lib") No config.

With thosepp\_lacked specifies the Driang masks defaulty the storage (recommended to the lack of the la The control of the co Vangred\_IL\_pipe indicates whether secondary-index WW magned\_IL\_pipe indicates whether secondary-index WW magned\_IL\_pipe indicates whether secondary-index WW magned\_IL\_pipe indicates in a secondary-index WW indicates or undefined\_IL\_pipe for IL.Q. (magned\_IL\_pipe, true), We Each of the following entries central has many Javascript We without machines are awailable for executing map, reduce, (Mappellar September 1), and control of the control of the (mappellar of the control of the co W js\_max\_vm\_mem is the maximum amount of memory, in megabytes, W allocated to the Javascript Whs. If unset, the default is ()s\_max\_vm\_mem, S), We je thread\_stack is the maximum amount of thread stack, in megabyes We allocate to the Javascript Wm. If unset, the default is 1875. We NOTE: This is not the same as the C thread stack. (6), thread\_stack, 16). applications of the second of (fun\_limit, 58000).
We object format controls which binary representation of a risk\_object.
We do not controls are used to the control of a risk\_object of the control of the control of a risk\_object.
We do risk of the control of th WR Risk Search Config (risk\_search, 'W To enable Search functionality set this 'true'. (enabled, false) We Size, in bytes, of the in-memory buffer. When this We threshold has been reached the data is transformed We into a segment file which resides on disk. (buffer\_pollower\_size, 1848376), (buffer\_rellower\_size, 1040570), %% Overtime the segment files need to be compacted. %% This is the maximum number of segments that will be %% quicker but more frequent compactions. \_imax\_compact\_segments, 20) Configure how Bitcask writes data to disk. erlang: Erlang's built-in file API nif: Direct calls to the POSIX C API Wi The NIT mode provides higher throughput for certain Www.konds, but has the potential to negatively impact With the Triang WM, leading to higher worst-case latencies Wi and possible throughput collapse. We elevelDB Config (elevelDb, [{dsta\_root, "./dsta/levelDb"} unfig.

What handlers to install with what arguments the files when We the derbuiks for the Lugfies are to retain the first when We the derbuiks for the Lugfies are to retain the first, and has the Lugfies for the first, and has the Lugfies for the first that the Lugfies for the foreign the derbuiks of the Lugfies for the first that the Lugfies for VE If you wish to have risk log messages to syslog, you can use a hand!
Vk (idger\_syslog\_backend, ["risk", daemon, info]),
((handlery f Vk Whether to write a crash log, and where.
Vk Commented/omitted/undefined means no crash logger
(crash\_log, "./log/crash.log"), We Maximum size in bytes of events in the crash log - defaults to 6553 (crash\_log\_esg\_size, 65536), We Maximum size of the crash log in bytes, before its rotated, set \(\frac{4}{4}\) to 8 to disable rotation - default is \(\frac{9}{4}\) (crash\_log\_size, 20482768), We What time to rotate the crash log - default is no time We rotation. See the lager README for a description of this format We https://githbo.com/basho/lager/blob/master/README.org (crash\_log\_date, "500") Vk Number of rotated crash logs to keep, 0 means keep only th Vk current one - default is 0 (crash\_log\_count, 3). We Configure the following : We of that event type. (busy\_port, true), (busy\_dist\_port, true) % SASL config (sast, [sast\_error\_logger, false) WA Authentication style used for access to the admin WA parel. Valid styles are 'userlist' < TOSOs. (auth, userlist), We If suth is set to 'userlist' then this is the NW list of usernames and passwords for access to the NW adding paged.

(userlist, [("user", "pass")]). Wi The admin panel is broken up into multiple Wi components, each of which is enabled or disabled Wh by one of these settings. (admin, frue)

## The enabled Yokcouns set this 'on'. yokcouns = off ## The port number which Solr binds to. vokcouns.solr port = 9899 ## The port number which Solr JMX binds to. yokozuna.solr\_jex\_port = 8985 promote the past to the foll AM. Non-standard ## Plan anguments, i.e. -Ci., may not be portable across 7M ## Suplementations. I.e. -Ci. overboompressed for force ## Suplementations. I.e. -Ci. overboompressed for force principles. I.e. - Justice - Justice - AM. visit force ## The data which to there all Vokunuse related data ## The data which to there all Vokunuse related data ## The data which to there all Vokunuse related data ## The data which to there all Vokunuse related data ## The data which to the data. ## leveldb data\_root leveldb.data\_root = ./data/leveldb Inself. Act, yes - (Intel/aced), 1 to Me of that which we have been presented by the control of leveldb.bloomfilter = on ## Set to 'off' to disable the admin panel. risk\_control = off ## Authentication style used for access to the admin ## panel. Valid styles are: off, userlist risk control auth = userlist \*\* If auth is set to 'userlist' then this is the \*\* list of usercases and passwords for access to the \*\* admin panel. we like of uncrease and passions for access to the risk, mortial, mark, mark, squared + pain we bitten data cost, marked + pain of the extraction of the cost of the cost of the extraction to bitten write, don't be extracted by the cost of the extraction of the extraction of the cost of the extraction of extraction of the extraction of ## Storage backend specifies the Erlang module defining the storage ## mechanism that will be used on this node. storage backend = bitcask we ray mum - Time

\*\*Retrict has fest AAK can build hash trees. Ruising the tree

\*\*for a given partition requires a full team over that partition's

\*\*Config is of the force:

\*\*Span-builds, par-timespan

\*\*station of the force:

\*\*Span-builds, par-timespan

\*\*station of the force:

\*\*station o we between the often tent tree are supline after thing builtse between the often tent tree are supline after thing builtse data stays consistent ofth the setual IV become data. It size
so data stays consistent ofth the setual IV become data. The size
set to be a supplied to the setual IV become data. The size
set infragance for performance reasons. The time is specified in
set infragance for performance reasons. The time is specified in
set instructions of the setual IV become the set of the setual IV become
set infragance for a large setual infrared to the setual IV become
set instructions of the setual IV become the setual IV become
set instructions of the setual IV become the setual IV becomes
set instructions of the setual IV becomes the setual IV becomes
set instructions of the setual IV becomes the setual IV becomes
set instructions of the setual IV becomes the setual IV becomes
set instructions of the setual IV becomes the setual IV becomes
set instructions of the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV becomes the setual IV becomes
set in the setual IV \*\* The tick determines have often the AMI manager looks for work ## to do (building/supiring trees, triggering exchanges, etc.) \*\* The default is every if seconds classing following such as the contract of the contract of the contract is a second of the contract of the anti\_entropy.tick = 15e \*\*B m directory where AMC hush trees are stared. anti\_entropy\_data\_dfr = ./data|anti\_entropy \*\*B ba Leaville options used by AMC to generate the LevellM-backs anti\_entropy\_antip\_antip\_time = 460 anti\_entropy.mus\_open\_files = 30 ew magned\_name is UNL used to subsit map/reduce requests to Risk mapred\_name = mapred \*\* mapred 21 pipe indicates whether secondary-index \*\* Royleduce imputs are quesed in parallel via their own \*\* pipe ('true'), or serially via a helper process \*\* ('false' or undefined). Set to 'false' or leave \*\* undefined during a rolling unreade from 1.8. ## Each of the following entries control how many lawsscript ## wirtual machines are available for executing map, reduce, ## pre- and post-commit hook functions. javascript, ws.mag.count =  $\bar{x}$ ## js may we see is the maximum amount of memory, in segubytes ## allocated to the lawscript WMs. If wrest, the default is ##  $940\,$ ## ji\_thread\_stack is the maximum amount of thread stack, in megaby ## allocate to the Javascript UM. If unset, the default is 1000. ## MOTE: This is not the same as the C thread stack. javascript\_vm.thread\_stack = 16 ## ju\_nource dir should point to a directory containing lawscript
## source files which will be leaded by Risk when it initializes
## lawscript NNs.
## jawscript NNs. ## passerugr\_m.sacc\_pr = rump/j\_carcae
## http://m.concing.detremises haw lisk trusts URL seconded
## buckers, bays, and links over the MCT #F. When set to car
el buckers, bays, and links over the MCT #F. When set to car
el Channels, pr. Sacc pr. ## Switch to wnode-based vclocks rather than client ids. This ## significantly reduces the number of vclock entries. ## Only set on if "all" nodes in the cluster are upgraded to 1.0 wnode vclocks = on \*\* This option toggles comparibility of keylisting with 1.8 \*\* and earlier versions. Once a rolling upgrade to a version \*\* > 1.8 is completed for a cluster, this should be set to \*\* then for better control of semony usage during key listing litties; backgressure - on. •• This option specifies how many of each type of fam may exist ex concurrently. This is for owerload protection and is a new ex mechanism that doublewest 1.2's health checks. Note that this number ef seprements two potential processes, so → in wn.args should be at ef least XX the fem\_limit. The speciment for protection of protection of the control of the c ## listemer.https.cnamer is an IP addre ## MITPS interface will bind. ## listemer.https.internal = 127.8.0.1: ## the number of replicas stored. Note: See CEP Controls for further discussion ## http://docs.hasho.com/risk/latest/dev/advanced/cap-controls/buckets.demplits.will. ## whether or not siblings are allowed. ## Note: See Wector Clocks for a discussion of sibling resolution buckets.default.last\_write\_wins = false ## Default ring creation size. Make sure it is a power of 2, ## e.g. 16, 12, 64, 128, 256, 512 etc ## rine size = 66 ## Default cert location for https can be overridde
## with the ski coefig variable, for example:
## ski.certfile = ./etc/cert.pen e\* Default key location for https can be overridden
e\* with the sul config variable, for example:
e\* sul.keyfile = ./etc/key.pen ## Default signing authority location for https can be overridden ## with the sil coeffg variable, for example: ## sil coeffile = /et/coeffile.or ew handoff.port is the TCP port that Risk uses for ew intra-cluster data handoff. handoff.port = \$800 \*\* To encrypt risk core intra-cluster data handoff traffif ew uncomment the following line and edit its path to an ew appropriate certfile and keyfile. (This example uses : ew single file with both items concatenated together.) ew bandoff sull-certfile = /tmg/ericaver.psm we mader all neverties - Transportaneous per ser Clean suggest and the Control of the Control o platform\_log\_dir + \_/log

\*\* where do you want the consule\_log camput:

\*\* off : Combone

\*\* file: the macfiled by log\_commole\_file

\*\* both : log\_commole\_file

\*\* both : log\_commole\_file

\*\* the thir | log\_commole\_file

\*\* the log\_lowlef of the commole\_log
log\_commole\_file

\*\* the log\_lowl of the commole\_log
log\_commole\_lowl = lof ## location of the console log log.console.file = ./log/console.log ## turn on sysleg log.sysleg = off ## Whether to write a crash log, and where. ## Commented/undited/undefined means no crash logger. log.crash.file = /log/crash.log ## Maximum size in bytes of events in the crash log - def log.crash.max size = 680 er Musimus size of the crash log in bytes, before its rotated, set set to 0 of disable rotation - default is 0 log.crash.size = 1892 ## Name of the risk node ## Cookie for distributed node communication. All nodes in the same cluste ## should use the same cookie or they will not be able to communicate. ## Increase number of concurrent ports/socket erlang.max.ports = 66000 ## Set the location of crash dumps erlang.crash\_dump = ./log/erl\_crash\_dump ew Raise the ETS table limit erlang.max ets tables = 256888 ## Raise the default erlang process limit erlang process limit = 256600 writing process limit = 500000

\*\*For mode with many buny first parts events, Busha recommends

\*\* raising the sender-tide method's distribution buffer time

\*\* 2000 may not be sufficient for ones everishment and is a suggested

\*\* starting point.

\*\*Est that it is about it is about it megabore).

\*\*Sea: that it is about it is about it megabore).

\*\*Sea: that it is about a suggested and it is a suggested and it is a suggested and it is a suggested.

\*\*Sea: that it is about it is about it is about it is about it is a suggested and it is a sugges ew Erlang UM scheduler tuning. ew Prenequicite: a patched UM from Rasho, or a UM compiled separate ew with this patch applied: ew https://pitc.github.com/evanecc/a500f4c6376388e4672e ew archive.de/.com/evanecc/a500f4c6376388e4672e em ering, that a see em Reading or writing objects bigger than this size will write em a warning in the logs. were object size a DMP ## Writing an object bigger than this will fail. max\_object\_time = \$890 ## Writing an object with more than this number of siblings ## will generate a warning in the logs. warn\_siblings = 25

ew Writing an object with more than this number of siblings ew will fail. max\_siblings = 100

#### Thanks